UNIVERSITE DU QUEBEC

MEMOIRE

PRESENTE A

L'UNIVERSITE DU QUEBEC A TROIS-RIVIERES

COMME EXIGENCE PARTIELLE

DE LA MAITRISE EN PSYCHOLOGIE

PAR

JOANNE CHARRIER

SUR LA METHODOLOGIE ET LA METROLOGIE DE L'OBSERVATION SYSTEMATIQUE

JUIN 1988

Table des matières

Sommaire

Depuis que le concept de comportement a été introduit en science, les procédures d'observation se sont élaborées avec de plus en plus de sophistication. De nombreux instruments de mesure du comportement ont été proposés selon la lecture du réel que les chercheurs ont tenté d'expliquer. Les préoccupations d'évaluation de ces filtres épistémiques sont devenues centrales et les méthodologies utilisées font l'objet d'une grande controverse. Celle-ci provient du fait que les définitions psychométriques classiques, historiquement formulées par rapport aux tests psychologiques, sont difficilement applicables aux méthodologies observationnelles contemporaines. La recherche actuelle se situe au niveau de cette problématique et vise la conceptualisation du processus observationnel sous un nouveau regard théorique, tout en proposant des modèles d'opérationnalisation de la fidélité et de la validité des données.

Le contexte théorique de la problématique concernée est élaboré autour d'une description de plusieurs stratégies observationnelles, d'une présentation des théories psychométriques classiques, d'une discussion sur les difficultés à utiliser les concepts traditionnels pour l'évaluation des données d'observation

naturelle, et d'un inventaire des multiples approches de fidélisation employées par les chercheurs étudiant le comportement humain.

Le processus observationnel est décrit d'un nouveau point de vue théorique par rapport aux composantes suivantes: phénomène que l'on veut circonscrire, rôle de l'observateur dans sa tâche d'observation, filtre observationnel. Finalement, des propositions d'opérationnalisation de la fidélité et de la validité sont formulées pour le cas d'un système d'observation des mouvements articulaires dans la communication. Ainsi, les principaux aspects qui sont développés concernent: (1) une méthodologie de comparaison des unités continues du codage multidimensionnel; (2) l'analyse avec critères de concordance entre les classements spatio-temporels des codeurs; (3) la conception d'une matrice présentant une description cinématique d'une séquence comportementale d'un sujet observé; (4) une procédure de validité des transformations spatiales d'un corps en mouvement, basée sur un argument géométrique.

Introduction

Dans le domaine de la recherche observationnelle, les problèmes liés à la gestion et à l'analyse des données d'observation ont été amplement commentés. Cependant, peu de recherches empiriques ont proposé des solutions aux questions méthodologiques concernant la valeur ou la qualité des données, si ce n'est pour adapter les théories traditionnelles de la fiabilité des mesures au contexte de l'observation directe du comportement. Plus souvent, les problèmes touchant la gestion des données ont été contournés en concevant des méthodes d'enregistrement et de codage qui contraignent les chercheurs à ne retenir qu'une faible partie de la réalité observée. A la limite, il serait justifié de se demander si l'ensemble des données comportementales publiées à ce jour reflète davantage le caractère de la "grille" utilisée par le chercheur que la réalité à l'étude.

Les objectifs de cette recherche sont, d'une part, de rendre compte des stratégies observationnelles les plus utilisées dans le contexte de l'observation naturelle et, d'autre part, de proposer une nouvelle approche conceptuelle de la fidélité assortie d'une stratégie d'étude de la fiabilité des données ainsi obtenues. Il s'agit d'utiliser le contexte d'une recherche portant sur l'élaboration d'une grille d'observation du comportement non verbal. Cette grille tient compte de la complexité de l'observation du mouvement humain; elle permet de relever, de façon continue et simultanée, les incidences de

mouvement dans plusieurs dimensions (espace, temps, amplitude). Les problèmes à résoudre ont trait à la définition d'unités quantifiables nécessaires à l'analyse statistique de la fidélité des données. Comment peut-on découper des données spatio-temporelles continues et simultanées pour que les classements de plusieurs observateurs ou du même observateur se prêtent à la comparaison? De plus, l'aspect de la validité des données peut être vérifié sous un angle méthodologique nouveau, compte tenu du caractère objectif et concret des observations produites avec cet instrument.

Ainsi, cette recherche tentera d'apporter des éléments de réponse quant aux concepts de fidélité et de validité faisant l'objet des préoccupations méthodologiques de la recherche observationnelle en sciences humaines. Aussi, des méthodes d'évaluation quantitative viendront compléter l'étude en vue d'opérationnaliser les nouveaux concepts.

La présentation de cette étude théorique suivra une démarche de façon telle que les propos abordés dans chaque chapitre serviront de toile de fond pour les questions élaborées au chapitre suivant. Le premier chapitre traitera des origines et des diverses stratégies observationnelles. Le deuxième chapitre présentera les concepts psychométriques classiques ainsi que le répertoire de techniques statistiques constamment employées par les chercheurs adeptes de l'observation; on y fera état de l'application controversée de ces concepts et techniques au niveau de la vérification des données provenant des méthodes d'observation naturelle. Enfin, le dernier

chapitre sera constitué des thèmes centraux de cette recherche: description d'une nouvelle méthodologie observationnelle, conceptualisation du processus observationnel, opérationnalisation de la fidélité et de la validité d'un instrument d'observation multidimensionnelle et continue.

Chapitre I

L'observation directe: ses origines, ses contextes d'utilisation et ses méthodes

Première partie: Ethologie et observation directe

L'étude des comportements par l'observation trouve principalement son origine dans le cadre conceptuel de l'éthologie. En éthologie contemporaine, les auteurs s'entendent pour distinguer deux principales orientations. Hinde (1982) parle de deux contextes: celui de la communication non verbale et celui de l'étude des relations interpersonnelles. Selon lui, l'origine de ces deux contextes provient du contact entre éthologistes et sociologues. Fassnacht (1982) mentionne simplement «éthologie animale», dite classique, et «éthologie humaine». Crook (1970: voir Trudel et Strayer, 1986) spécifie les intérêts de chaque orientation comme suit:

> L'éthologie dite classique centre surtout l'attention sur l'analyse détaillée des schèmes comportementaux, de leurs antécédents immédiats et de leurs développements subséquents. Par contraste, l'éthologie sociale s'intéresse davantage aux questions d'ordre fonctionnel en regard, par exemple, de la coordination des schèmes d'échanges dyadiques, de l'organisation des relations sociales parmi les individus, et des variations écologiques dans la nature des structures de groupe (p.169).

Vauclair (1984) parle de l'observation comme d'une "caractéristique

essentielle de ce que l'on peut appeler l'attitude éthologique"(p.123). Il ajoute que l'observation désigne "tout recueil de données établi à partir de la description du comportement spontané des animaux dans leur milieu naturel" (p. 123). Hinde (1966: voir Vauclair, 1984) définit deux types de descriptions éthologiques:

> Le type empirique se réfère à une description des changements spatio-temporels dans les mouvements des membres et du corps.(...) Le second type est appelé description fonctionnelle dans la mesure où le comportement est incorporé à un cadre de rérérence qui définit sa fonction proximale ou ultime (p. 124).

Le premier type est une description libre d'inférence constituant des catégories de comportements plus proches de la réalité observée. Par contre, une description fonctionnelle suppose un degré d'abstraction plus élevé entraînant la combinaison des comportements par catégories génériques. L'inconvénient réside dans le danger de faire des erreurs d'interprétation.

Les questions fondamentales en éthologie

Fassnacht (1982) présente une classification des questions fondamentales en éthologie. En premier, les questions phylogénétiques, constituant un aspect de la recherche comparative sur le comportement, se rattachent aux concepts d'homologie ou de morphologie comparée. Ces questions appartiennent plus exclusivement à l'éthologie classique, i.e.

animale; la découverte des homologies dans les comportements d'instinct est moins complexe que dans les comportements humains. Il y a peu d'études empiriques qui font des comparaisons directes entre l'humain et l'animal sauf celles qui viennent de la psychologie développementale. Les comparaisons des répertoires comportementaux des jeunes enfants avec ceux des jeunes animaux sont les plus représentatives.

Les trois autres domaines, fonction, causalité et ontogénie du comportement, comportent des objectifs existant aussi indépendamment de l'éthologie. En éthologie, ils sont abordés sous l'angle de l'orientation biologique. La question de fonction a deux acceptions: soit la valeur de survie d'un comportement particulier pour l'individu et pour l'espèce, soit l'effet de ce comportement sur l'individu et sur son environnement physique et social. L'approche fonctionnelle des études sur le comportement humain est représentée dans le domaine de la psychopathologie, où elle aide à expliquer le comportement pathologique; d'autres études au niveau du développement de l'enfant font ressortir la fonction biologique de certains patrons de comportements (Fassnacht, 1982).

La question de causalité est reliée à la découverte des facteurs physiologiques orientant certains schèmes comportementaux. Cet objectif est aussi en liaison avec le domaine de la psychologie expérimentale. L'investigation causale s'étend à plusieurs champs de recherche et certains éthologistes parlent même de «physiologie béhaviorale» (Fassnacht, 1982). Finalement, la question d'ontogénie, relevant du domaine de la psychologie

développementale, porte, en éthologie empirique, sur le développement intra-espèce: Comment l'individu s'est-il développé pour devenir ce qu'il est? Une priorité est accordée au but adaptatif du comportement en fonction d'une orientation biologique. Les recherches sur les relations et les interactions «mère-enfant» sont quelquefois associées à ce domaine de l'éthologie.

Une cinquième perspective dans l'analyse du comportement en éthologie est identifiée par Kummer (1971: voir Trudel et Strayer, 1986); elle concerne l'organisation même du comportement, c'est-à-dire sa structure. C'est au niveau de l'analyse structurelle que l'éthologie sociale place une emphase particulière. Trudel et Strayer (1986) résument cette démarche comme suit:

> la description détaillée des divers niveaux d'organisation du comportement constitue la base empirique nécessaire à l'identification des schèmes comportementaux, à l'étude de ces structures dans une perspective causale et fonctionnelle, à l'analyse des modifications qualitatives et quantitatives de ces activités au travers du développement, ainsi qu'à l'examen de l'évolution des comportements, des relations et des structures de groupe (p. 170).

Les phases méthodologiques en éthologie

Il existe plusieurs méthodologies de recherche en éthologie; celle de l'observation directe du comportement constituera le propos de cette section. Un principe de base en éthologie stipule que le comportement devrait être étudié dans son environnement naturel avec le moins de perturbations possible.

La méthode d'observation directe devient donc toute désignée. Fassnacht (1982) distingue trois phases dans cette démarche. La phase initiale dite d'observation exploratoire consiste principalement à connaître le comportement de l'espèce choisie. A ce stade de familiarisation, l'éthologiste tente de représenter le flot comportemental et d'en décrire les unités dans un langage relativement non technique. En éthologie humaine, cette phase est souvent plus restreinte qu'en éthologie animale. L'inventaire comportemental élaboré à partir de cette connaissance de l'espèce est appelé un éthogramme; sa construction représente la deuxième phase de la démarche. Cet éthogramme présente une description physique du comportement ou une description par conséquence (Hinde, 1959: voir Fassnacht, 1982). L'idéal d'une description complète est rarement atteint pour le comportement animal et il est encore moins probable dans la formation d'un éthogramme du comportement humain.

La phase d'observation systématique devient possible une fois l'éthogramme bâti. Cette période d'observation se déroule selon des règles définies concernant entre autres le comportement-cible, la période d'observation, le temps d'observation, sa durée, la question à quoi l'on veut répondre, les individus à l'étude et la technique d'observation. Les questions primordiales de l'éthologie ont toutes leur signification dans ce contexte. L'observation systématique permet de trouver des homologies dans les patrons de comportement et donne donc le pas à l'approche comparative. Pour les éthologistes, les investigations empiriques réalisées à cette étape constituent les préliminaires nécessaires à toute expérimentation mettant l'emphase sur la

causalité d'un phénomène. Vauclair (1984) déclare que l'éthologie d'aujourd'hui fait un usage égal des manipulations expérimentales et de l'observation naturaliste. Ces deux approches se justifient dans l'atteinte des objectifs propres à l'éthologie. Elles permettent de résoudre les problématiques concernant "la nature des êtres vivants et leurs interactions avec l'environnement, le rôle explicatif du comportement dans l'analyse de ces interactions et enfin la signification évolutive des modifications du comportement étudié" (Vauclair, 1984; p. 134). Vauclair identifie trois techniques de sélection des observations auxquelles les éthologistes ont recours: (1) l'échantillonnage centré sur un individu-cible, (2) l'échantillonnage de séquences d'interactions et non pas sur un individu particulier, et (3) l'échantillonnage instantané ou à des moments prédéterminés dans le temps.

La méthode d'observation systématique a aussi une longue histoire en psychologie (Longabaugh, 1980; Beaugrand, 1982). La section suivante expliquera les étapes d'une telle démarche systématique lorsqu'appliquées dans le contexte d'observation du comportement humain. En psychologie, le domaine de recherche sur le comportement non verbal dans la communication humaine reconnaît l'apport important des données empiriques obtenues par la rigueur des méthodes éthologistes. Les propos de Cosnier et Brossard (1984), cités ci-après, en donnent témoignage:

> La multicanalité de la communication humaine avait certes été reconnue depuis longtemps. (...) mais c'est à l'époque contemporaine que la conception de la communication multicanale a

> été élargie, précisée et étayée par les réflexions et les travaux des éthologues (...).
>
> La familiarité des éthologues (...) avec les systèmes de communication animale les prémunit en effet contre le préjugé traditionnel de l'exclusivité du canal acoustique (...). Cette approche appliquée à l'espèce humaine a permis de concevoir aujourd'hui que le caractère acoustique n'était pas un critère nécessaire pour définir le langage, mais que le critère fondamental résidait plutôt dans l'existence d'un lien conventionnel entre les signaux et leurs référents. (...) Ainsi les conceptions éthologistes prédisposent à s'apercevoir que la communication langagière déborde largement le seul système verbal.

En résumé, deux contributions majeures peuvent être attribuées aux méthodes éthologiques. D'abord, elles fournissent des méthodes d'observation et d'analyse du comportement; ensuite, elles procurent des hypothèses pouvant être vérifiées par les études directes de l'être humain (Barnett, 1980).

Cette première partie a montré la valeur de la méthode d'observation directe dans une démarche de compréhension des comportements d'une espèce quelconque. Un fait demeure: c'est que la méthode d'observation directe occupe une place unique en psychologie. Des préoccupations d'ordre biologique à propos de la valeur adaptative et de la signification évolutive des comportements obligent à la définition de méthodes d'observation systématiques. Plusieurs domaines de recherche en psychologie appliquent les méthodes observationnelles. Une énumération des principaux domaines est donnée par

Beaugrand (1982; p. 167): la recherche sur le développement cognitif et social, sur l'évaluation psychopathologique, sur l'apprentissage animal et sur l'étude des comportements accompagnant une réponse instrumentale. Ce dernier champ d'intérêt rejoint le contexte de notre recherche, lequel s'appuie sur un système d'observation de l'aspect non verbal de la communication ou des interactions humaines. Afin de mieux distinguer les méthodes observationnelles dans la recherche en psychologie, nous consacrerons la prochaine section à décrire les composantes d'un plan d'observation systématique.

Deuxième partie: Description du processus observationnel

Lorsque l'observateur poursuit un but systématique, il observe de façon sélective les comportements dans lesquels il reconnaît les «concepts-clés» mis à jour au cours des phases précédentes d'exploration et de description des comportements-cibles prélevés dans leur environnement naturel (Longabaugh, 1980). Bunge (1984) définit l'observation comme "une perception «préméditée» parce qu'elle est faite dans un but bien défini et «éclairée» parce qu'elle est guidée par un corps de connaissances" (p. 47). Reprenant la nomenclature simple de cet auteur, c'est-à-dire que "l'objet de l'observation est un fait actualisé" et que "l'issue d'un acte d'observation est une donnée", il est plus facile d'entrevoir comment le processus observationnel est amorcé. La pertinence et la qualité de ces deux aspects seront déterminés par le plan

d'observation. Celui-ci se compose de quatre principales étapes représentant les tâches majeures impliquées dans le processus observationnel. La première étape consiste à décrire le comportement à étudier; elle est donc fondamentale pour l'orientation du plan d'observation. En second lieu, le plan d'observation vise à définir une méthodologie de cueillette des données qui est en relation directe avec les objectifs de recherche. En troisième lieu, il permet de prévoir la mise en fichier des données brutes d'observation. C'est l'étape de la gestion des données. Et finalement, il oriente sur la méthode d'analyse des données, étape au cours de laquelle les données brutes seront transformées dans une forme statistiquement analysable. Nous garderons cette même chronologie pour expliquer comment s'élabore un plan d'observation.

Elaboration du plan d'observation

A. Description du comportement à l'étude et choix d'une taxonomie

L'élaboration du plan d'observation débute par la description du comportement que l'on veut étudier. Un concept aussi vaste que le comportement oblige, dans le domaine de l'observation, à l'identification des propriétés pertinentes sur lesquelles s'appuie l'observateur pour reconnaître de manière fidèle des unités comportementales. Beaugrand (1982) identifie toute une gamme d'unités, passant des descriptions moléculaires jusqu'aux plus molaires. Par exemple, lorsque l'objet d'étude concerne les composantes motrices des schèmes moteurs, nous avons une définition plus moléculaire que

pour celui ayant trait aux interactions sociales entre deux et plusieurs individus.

De plus, Beaugrand (1982) définit deux ensembles de critères de classification des unités comportementales: les critères concrets classifient les comportements selon leurs propriétés formelles, topologiques ou cinétiques, ou encore selon leurs effets sur l'environnement; les critères théoriques ou abstraits donnent des interprétations causales ou fonctionnelles de certains indices comportementaux concrets. Dans le même sens, Longabaugh (1980) parle plutôt des critères du «quoi» et du «comment»; les premiers visent à établir une classification sur le contenu du comportement et les seconds sur le contexte, c'est-à-dire sur l'explication de la façon dont le contenu est exprimé.

Toutefois, les niveaux d'abstraction des comportements étudiés relèvent de l'approche du chercheur et cela, sans égard aux critères de classification choisis. Ainsi, une approche éthologiste limite ses observations à des unités comportementales moléculaires dans le but de constituer un éthogramme, alors qu'une approche écologiste englobe des unités molaires informant sur le contexte social du comportement et sur l'interaction des acteurs. Enfin, d'autres conditions sont à respecter dans l'identification des unités à des critères concrets ou abstraits; elles concernent entre autres la qualité, la quantité, l'objectivité, l'homogénéité, la spécificité et la complexité des unités comportementales choisies (Beaugrand, 1982; Hollenbeck, 1978).

Vient ensuite le moment de choisir et de définir les unités comportementales retenant l'attention du chercheur. Des règles générales s'appliquent: le chercheur ne doit retenir que les unités directement associées aux objectifs de la recherche et leur conférer une étiquette descriptive et succinte. Cette description inclut un critère de référence qui la rend opérationnelle et directement observable. Son but est de prévenir les interprétations et les variations entre observateurs, ou chez le même observateur. Hollenbeck (1978) explique que la définition de l'unité d'observation a pour but de contrôler plusieurs aspects du processus observationnel comme les comportements observés, les catégories de codage utilisées et l'interaction entre l'observateur et le système de codage. Longabaugh (1980) ajoute que les spécifications de l'unité d'observation concernent les acteurs (émetteurs, récepteurs, individus, groupe, etc.), les cibles des comportements (identifiées ou pas) et les sites (temporels, spatiaux, culturels, restreints ou non).

Ainsi, le chercheur doit obtenir une taxonomie spécifique aux objectifs de sa recherche. A partir des unités comportementales qu'il a choisies et définies, il élabore le répertoire des comportements qui seront observés et qui devront être reconnus pour la notation ou l'encodage. L'étendue de ce répertoire, c'est-à-dire de la taxonomie construite par le chercheur, représente, comme nous l'avons expliqué, le niveau plus ou moins profond d'encodage qui sera appliqué à l'objet d'étude. Par exemple, une taxonomie de surface peut être constituée d'unités décrivant des patrons de comportement

spécifiques chez un groupe particulier de sujets (Ex.: taxonomie du comportement ludique dans le groupe chez des enfants de pré-maternelle), alors qu'une taxonomie plus profonde empruntera une notation plus molaire (Ex.: la catégorie «comportement agressif»). Pour Sackett et ses collaborateurs (1978), cette distinction se fait au niveau des unités décrivant des réponses motrices et celles décrivant des réponses vocales. Alors une taxonomie moléculaire emprunte des unités très rapprochées des actions motrices spécifiques, ou des postures, des gestes, des expressions faciales, des objets et des directions de l'action. Au contraire, une taxonomie molaire utilise des classes de combinaisons d'un certain nombre d'actions, de directions et d'objets de comportement se définissant par la fonction ou le résultat des actions motrices. Les taxonomies du comportement servent donc à abstraire des unités de mesure à partir des actions des sujets. Des critères concernant en premier l'efficacité de la recherche en fonction des hypothèses énoncées et ensuite, la sensibilité des mesures anticipées orienteront le chercheur dans le choix d'une taxonomie appropriée (Beaugrand, 1982). Enfin, la taxonomie choisie renseignera sur quoi observer et déterminera par conséquent qui, où, quand et comment observer.

Les questions du «qui», «où», «quand» et «comment» observer montrent que le choix d'une taxonomie est lié de très près aux modalités d'enregistrement des unités comportementales, donc aux étapes de la cueillette et de la gestion des données. Aussi, la section suivante enchaîne sur les considérations propres à la façon d'enregistrer le comportement à l'étude. Bien

que dans l'élaboration du plan d'observation ces deux entités sont considérées simultanément, et que quelquefois les chercheurs n'en font pas la distinction, nous tenons à les décrire de façon séparée pour faire ressortir les préoccupations distinctes qu'elles comportent. Nos propos porteront en premier sur les stratégies d'enregistrement à l'étape de la cueillette des données. Une autre section suivra avec les méthodes de codage utilisées à l'étape de la gestion des données.

B. Etape de la collection des données

L'enregistrement du comportement comporte des décisions préalables concernant les critères de répétition, de transition (fréquences), de ponctuation (état, événement, activité) et d'interruption des unités comportementales Ces critères donnent lieu à la systématisation de la collection des données et diverses stratégies d'enregistrement peuvent être envisagées. La planification de ces stratégies entourera les questions suivantes: "Le comportement à observer sera-t-il défini en événements ponctuels ou en événements avec une durée significative?" "Le contenu d'enregistrement informera-t-il seulement sur la présence du comportement-cible, ou aussi sur d'autres dimensions telles que la durée, avec ou sans notation des débuts et des fins d'occurrence, ou encore les séquences, l'ordre d'apparition, l'origine, la destination ou l'orientation du comportement-cible?" "Les périodes d'observation seront-elles faites sur le vif ou en différé, de façon continue et complète, ou par échantillonnage?" "La planification des périodes

d'échantillonnage se fera-t-elle aléatoirement ou selon l'occurrence du comportement-cible, ou encore selon des intervalles de temps pré-déterminés?" "Le comportement enregistré aura-t-il à être traité en termes de fréquence ou de proportion, ou en plus avec sa durée, son intensité et sa périodicité?" La résolution de ces interrogations amène le chercheur à concevoir simultanément ses stratégies reliées à la collection et à la gestion des données. Il doit choisir un système d'enregistrement des comportements tenant compte des dimensions qu'il désire enregistrer, lequel, par ailleurs, permettra d'échantillonner les données comportementales selon les techniques satisfaisant aux objectifs de la recherche.

A ce stade du processus observationnel, la terminologie employée prête quelquefois à l'équivoque. De plus, les sytèmes d'enregistrement sont fréquemment confondus avec les méthodes de codage; certains parlent de systèmes ou de procédures de codage (Frey et Pool, 1976; Longabaugh, 1980), de méthodes de quantification des données (Fassnacht, 1982), et enfin d'autres (Altmann, 1974; Sackett, 1978; Beaugrand, 1982) font appel aux techniques ou méthodes d'échantillonnage du comportement. Toutefois, les propos qui suivent concernent uniquement les systèmes d'enregistrement; ce n'est qu'à la section de la gestion des données que nous décrirons les méthodes d'échantillonnage.

1. Les systèmes d'enregistrement des observations selon diverses classifications. Fassnacht (1982) caractérise un système d'enregistement selon deux aspects: (1) la similitude ou la différence des unités descriptives; (2) la relation particulière entre ces unités. Il distingue quatre catégories de

systèmes d'enregistrement, soit les systèmes de types verbal, nominal, dimensionnel et structural. Ces catégories se recoupent mutuellement. L'auteur attire l'attention sur le fait qu'un système d'enregistrement produit des opérations se retrouvant à un niveau de représentation tertiaire et que les opérations ne deviennent significatives qu'une fois replacées au second niveau, soit celui des perceptions. Dans bien des cas, le système de représentation observationnelle n'est pas isomorphique avec le système perceptuel utilisé. Un ensemble de relations bien spécifiques entre les unités est introduit par le système d'observation indépendamment des objets décrits, et le danger de fausses conclusions sur l'objet d'observation est grand[1]. Pour Fassnacht, un mécanisme de représentation représente correctement lorsque les mêmes segments de réalité sont toujours représentés de façon identique. C'est à ce niveau que le chercheur doit avoir recours à l'algèbre et à la psychométrie dans le but de transformer des perceptions en codes pouvant être analysés.

Parmi les systèmes de type verbal, il cite (1) les notations sous forme de journal - méthode utilisée en premier par Darwin -, (2) les notations sous forme de spécimen développées dans une orientation écologiste, (3) la technique de l'«anecdote» ("critical incident") utilisée dans la classification des symptômes cliniques en psychologie, et enfin, (4) l'échantillonnage d'un événement comportemental unique dans sa forme de description verbale. Ces systèmes d'enregistrement ont en commun la représentation verbale du

---

[1] Laurencelle (1986) discute aussi des biais introduits par le système perceptuel de l'observateur: il distingue «processus de réalité» et «processus d'observation» en les définissant respectivement comme la portion du réel à repérer par les mesures et la description codée de cette portion.

comportement dans un langage de tous les jours. Ces données verbales doivent donc être assujetties à une décomposition de contenu avant que puissent s'effectuer les analyses mathématiques concrètes. Ces systèmes exposent les données à une deuxième possibilité de perte d'information, soit la transformation du récit de l'observateur par l'analyste.

Les systèmes de type nominal utilisent aussi une représentation verbale mais avec une proportion moindre de langage structuré. Fassnacht en distingue deux par la façon dont les unités des systèmes sont mises en relation: les systèmes utilisant des étiquettes caractéristiques et ceux employant des catégories[1]. Ainsi, les systèmes à catégories contiennent des unités closes et incompatibles entre elles nécessitant un seul enregistrement des unités dont la totalité enregistrée égale la longueur de la séance et décrit complètement un aspect du comportement. Par contre, les sytèmes avec étiquettes n'ont pas de critères rigides quant à la compatibilité des unités, ni sur la simultanéité des représentations entraînant un temps d'observation non équivalent au temps total des unités enregistrées, ainsi qu'une description incomplète du comportement.

Les systèmes de type dimensionnel impliquent des descriptions comportementales selon une relation graduée où les unités mutuellement exclusives prennent une valeur quantitative. Ces sytèmes d'enregistrement

---

[1] Ces deux classe de systèmes de type nominal correspondent aux catégories de systèmes que Laurencelle (1986) identifie comme les «listes de codes représentant des événements ponctuels» (étiquettes caractéristiques) et «l'enregistrement continu d'un comportement-cible (catégories).

font référence aux échelles de mesure (échelles nominale, ordinale, d'intervalles, de proportions) largement utilisées en psychologie et sur lesquelles plusieurs types de statistiques peuvent être appliqués. L'auteur rappelle le danger, surtout présent avec ces systèmes, de confondre les signes conventionnels opérationnalisables du système tertiaire avec une réalité perceptuelle qui n'est pas du même ordre. Il réfère à Hutt et Hutt (1974) pour distinguer les quatre principales méthodes de quantification du comportement qui sont effectuées avec les sytèmes de type dimensionnel. Ce sont les descriptions selon la fréquence, la durée, l'intensité et selon le comportement global, ou encore selon des combinaisons de ces aspects. Cependant, tous les concepts psychologiques ne se prêtent pas facilement à des évaluations quantitatives sur toutes ces dimensions, et particulièrement au niveau de l'intensité et de la globalité d'un comportement. Par exemple, la déclaration «Robert est plus agressif que Jacques», après observation et quantification de certaines unités comportementales définies en termes d'agression, reflète davantage une interprétation sur les intentions de ces individus plutôt que la mesure objective de leurs comportements. Le concept d'agressivité, pris ici dans son sens global, est difficilement opérationnalisable sans inférence ou biais majeur.

Finalement, Fassnacht identifie les systèmes de type structural comme une extension de ceux du type dimensionnel à cause de l'ajout des construits spatiaux dans la description du comportement. Une perception devient configurée ou hiérarchisée plutôt que d'être une dimension purement linéaire.

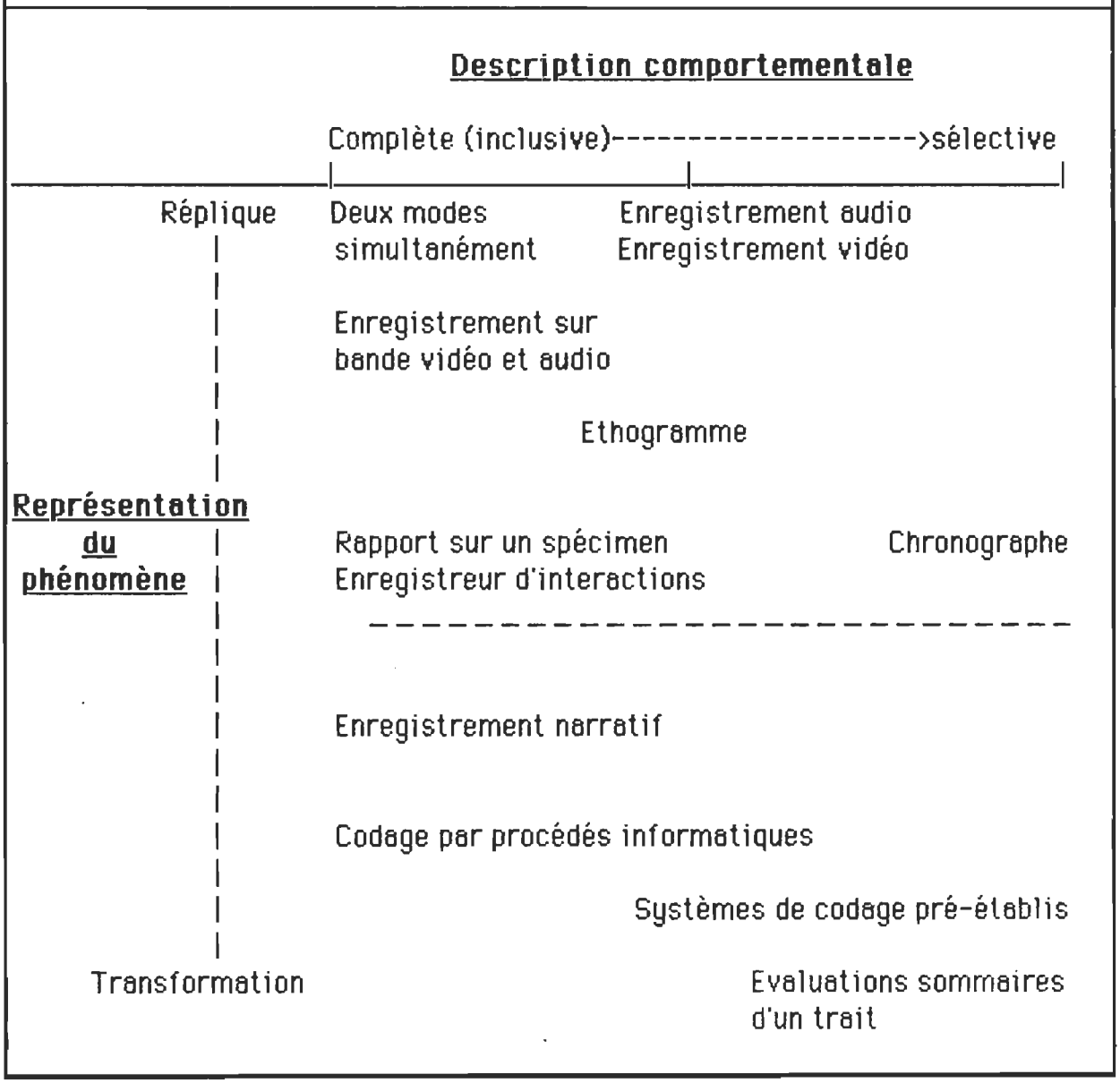
Ces systèmes ressemblent grandement aux précédents par le fait que l'objet «réel» est connu par une ou plusieurs positions dans un système de relations idéalisé. Ce type de systèmes est peu utilisé en raison des difficultés à reproduire directement à partir des unités un modèle comportemental précis.

Longabaugh (1980) apporte une classification des systèmes d'enregistrement assez différente de celle de Fassnacht puisqu'elle est basée sur deux autres dimensions combinées. La première se situe sur le continuum «complet/sélectif» qui est prévu dans l'enregistrement au niveau de la description du comportement; la deuxième concerne le degré de «représentativité» du phénomène étudié, passant de la simple réplique à la transformation en un index de construits théoriques. L'auteur organise un schéma bidimensionnel et obtient un tableau quadruple représentant une classification combinée des différents systèmes d'enregistrement. Ce tableau est reproduit sur la page suivante pour faciliter l'accès conceptuel à ce matériel.

Ainsi, on peut observer que les enregistrements audio et vidéo fournissent une réplique totale de l'événement original. De plus, si les deux modes sont employés simultanément, on obtient une description comportementale plus complète ou presque idéale, en admettant cependant qu'un rapport comportemental complet sur un organisme n'est pas possible. Par contre, aux autres extrémités du continuum, les sytèmes ont pour but de mesurer des construits théoriques et le matériel codé ne peut être retranscrit dans une réplique du phénomène. Les systèmes de type dimensionnel de

Tableau 1

Classification des systèmes d'enregistrement en fonction de la complétude de la description comportementale et du degré de transformation du phénomène



Fassnacht se retrouvent dans cette section. Ce tableau décrit bien combien la distance entre le phénomène comportemental et le construit théorique entraîne de transformation durant l'enregistrement des données. De façon similaire, l'axe concernant la description du comportement montre comment les méthodes d'enregistrement réduisent leur «prise de vue» à mesure qu'elles regroupent ou sélectionnent les comportements.

Plusieurs auteurs (Altmann, 1974; Bakeman, 1978; Holm, 1978; Sackett, 1978; Beaugrand, 1982) identifient simplement deux classes de systèmes d'enregistrement. La première classe regroupe toutes les méthodes d'observation continue et complète par lesquelles on vise à obtenir un récit descriptif du phénomène étudié. Avec cette stratégie, le chercheur enregistre tous les comportements d'intérêt dans le temps continu réel. De nos jours, les outils technologiques, tels les appareils magnétoscopiques ou audiophoniques, remplacent de plus en plus les méthodes papier-crayon, tels le journal de bord, les listes de contrôle, les matrices, car ils facilitent la saisie objective des phénomènes sans intrusion majeure. La deuxième classe contient les méthodes utilisant un échantillonnage du temps d'observation. Les systèmes de cette catégorie sont conçus pour enregistrer des intervalles de durée, de fréquence, de rythme. L'échantillonnage des comportements-cibles peut se faire à base temporelle régulière, le plus souvent des intervalles de 5, 10, ou 15 secondes, ou irrégulière (intervalles successifs avec fréquence modifiée); ce peut être aussi un échantillonnage déclenché par des événements-cibles(lorsque tel comportement ou tel ensemble de comportements apparaît ou n'apparaît pas),

ou finalement déclenché par les circonstances ou l'environnement (tel milieu plutôt que tel autre, telle période du jour, etc.).

Pour compléter cette section, une description des différents systèmes d'enregistrement souvent mentionnés dans la littérature de recherche nous fera découvrir que les façons de procéder sont multiples et qu'elles ne sont limitées que par des préoccupations d'efficacité ou de coût, ou encore par la créativité des chercheurs.

2. Les systèmes d'enregistrement selon leurs caractérisations. Holm (1978) décrit cinq techniques plus courantes:

a. Enregistrement de la voix. C'est un récit narratif sur ce qui se passe durant la séance d'observation. Cette méthode s'utilise mieux pour une recherche exploratoire, où l'on tente de cerner un phénomène particulier. Quelquefois, on enregistre seulement la fréquence et la durée totale des comportements visés; dans certaines recherches, on veut à la fois préserver l'occurrence et la durée selon la séquence d'origine. Le principal désavantage de ces méthodes est relié à l'effort requis par la suite pour déchiffrer l'enregistrement et le rendre utilisable pour les analyses ultérieures. De plus, cette transformation expose les données à de la manipulation interprétative et donc, à une correspondance affaiblie avec l'objet initial d'observation.

b. Enregistreurs d'événements. Ce sont des appareils pré-programmés permettant de mémoriser des catégories spécifiques de comportement (20 au plus). L'observateur appuie sur le bouton de la catégorie appropriée lorsque le

comportement se produit. Et si l'appareil est muni d'un lecteur de temps, la durée sera aussi enregistrée selon la période pour laquelle le bouton est maintenu enfoncé. Les limites sont évidentes: une seule catégorie à la fois, codage réductionniste, peu de catégories différentes, difficulté d'obtenir la fréquence réelle des comportements, etc. Ces méthodes ne conviennent qu'avec des buts très spécifiques, telle la description molaire de certains aspects des comportements d'un individu ou d'un groupe (Ex.: évaluation des manifestations d'intérêt cognitif chez un étudiant à sa classe de Maths). Avec ces méthodes aussi, il est impossible de reconstituer la période d'observation de l'objet d'étude telle qu'elle s'est produite.

c. Listes de contrôle ("checklists"). Ce sont des listes plus ou moins élaborées classifiant les comportements-cibles sous une étiquette ou sous une catégorie. Dans une forme plus sophistiquée, elles divisent la période d'observation en blocs de temps (ou intervalles temporels). Toutefois, même en limitant l'enregistrement à une occurrence par intervalle avec, en plus, des intervalles très petits et en restreignant le nombre de catégories pour obtenir le meilleur seuil de fiabilité des observateurs, on ne peut vérifier les mesures réelles de temps, de fréquences et de séquence de données.

d. Chronomètres et compteurs. Lorsque le nombre de catégories est petit et que l'information séquentielle n'est pas requise, ces techniques d'appoint permettent d'enregistrer avec plus de fiabilité les aspects temporels et de fréquence du phénomène observé. Ces raffinements, qui en fait se juxtaposent aux méthodes précédentes, assurent de meilleures utilisations et

généralisations des données d'observation, sans toutefois favoriser une reproduction «intacte» du phénomène observé.

e. Claviers à affichage numérique. Ce sont des enregistreurs de données sur un médium lisible par ordinateur. Les catégories de comportement à observer sont représentées par des codes numériques uniques. Ces systèmes de codes donnent accès au codage hiérarchique des comportements.

Beaugrand (1982) et Longabaugh (1980) mentionnent que les techniques audio-visuelles, comme les enregistreurs à cassettes audiophoniques ou magnétoscopiques, aident à l'encodage et à l'enregistrement des observations. Longabaugh reconnaît que l'investigateur doit maximiser les capacités visuelles et auditives de l'observateur. Ainsi, les mécanismes d'appoint comme la caméra, le microphone, ou autres enregistreurs électroniques, sont tout à fait indispensables à une réplique fidèle du phénomène à décoder. Les données enregistrées de cette façon se conservent avec un minimum de distorsion et demeurent disponibles pour de multiples investigations.

Enfin, Beaugrand et Longabaugh décrivent un dernier type d'aide instrumentale à l'enregistrement des données; il s'agit de l'éthographe, un appareil conçu pour produire des listes de codes pouvant être transmises à un ordinateur. L'éthographe contient un microprocesseur servant à enregistrer les données par codes à partir d'un clavier muni d'un certain nombre de touches et de boutons à levier. Les versions les plus connues sont le «Datamyte» (900 et 1000) et le «More» (OS-II et OS-III). Avec un éthographe, l'encodage peut

s'effectuer en temps réel, en tenant compte de la durée de l'événement; il permet d'encoder les données sur le vif ou à partir d'un enregistrement magnétoscopique. En plus de faciliter l'encodage de données quantitatives et qualitatives, ces appareils aident à enrichir la gamme des renseignements tirés des observations et à en dégager plus rapidement les analyses. Ils donnent accès aux traitements informatiques et donc à une meilleure vérification statistique des informations recueillies.

Les systèmes d'enregistrement sont certes des moyens visant à systématiser la collection des données. Cependant, ils n'assurent pas à eux seuls la constance et la fiabilité des observations. L'observateur humain effectue encore, malgré tout le dispositif en place, la majeure partie de la discrimination complexe du phénomène à l'étude. Toutefois, ses récepteurs perceptuels et ses habiletés à juger peuvent limiter ses observations. Sackett (1978) spécifie que les méthodes d'échantillonnage des comportements durant l'observation sont des solutions aux problèmes que posent les limites de l'observateur. Plusieurs de ces méthodes seront ci-après relatées.

C. Etape de la gestion des données

Cette étape consiste à transformer le comportement en unités spécifiques auxquelles on assigne des propriétés. Ce sont en fait les variables auxquelles le chercheur s'intéresse. Une abstraction sélective du comportement à l'étude résulte de cette transformation. Il convient donc de bien connaître les diverses méthodes de codage ou d'échantillonnage en usage

dans la recherche observationnelle, puisqu'une meilleure connaissance du filtrage qu'elles imposent aux données comportementales aidera à concevoir la nécessité d'appliquer aux données d'observation des méthodes d'évaluation pertinentes et rigoureuses.

Avant de passer à la description de plusieurs méthodes spécifiques d'échantillonnage, nous voulons distinguer deux principales catégorisations que les auteurs (Sackett, 1978, Longabaugh, 1980 et Fassnacht, 1982) identifient comme suit: l'«échantillonnage de temps» et l'«échantillonnage d'événements». L'échantillonnage du temps est décrit par Fassnacht (1982) comme une procédure entraînant des mesures approximatives de la durée et de la fréquence du comportement car elle impose à une séquence de comportements, une grille de courts intervalles standardisés. L'occurrence de comportements prédéfinis est enregistrée par l'observateur dans une succession d'intervalles de temps sur la base du «tout ou rien». Des intervalles plus courts sont choisis lorsque le but est d'atteindre le plus grand degré possible de précision temporelle. Toutefois, s'il y a autant d'intervalles que de comportements observés, la tâche sera plus laborieuse pour l'observateur et la procédure n'en sera pas plus représentative de la séquence comportementale originale. Longabaugh (1980) juge que les distorsions résultant de ce type de découpage font qu'il n'est point recommandable. En alternative, il suggère des intervalles de temps plus courts que le temps d'exécution du comportement et l'observation d'un seul attribut comportemental, ou encore, le codage par intervalles de fréquence modifiés. Sackett (1978) énumère quelques avantages et désavantages de

l'échantillonnage du temps. Les premiers sont: (1) le minimum de coûts, temps et efforts requis de la part du dispositif et de l'observateur, ainsi que de son entraînement; (2) l'assurance de fidélité par des codes bien définis. Les désavantages se situent par rapport à l'insensibilité à détecter des comportements peu fréquents et à différencier les comportements à fréquence modérée de ceux dont la fréquence est plus élevée; un autre inconvénient est la difficulté d'interpréter directement les scores de fréquence ou de durée des comportements puisqu'ils n'ont pas de vraie unité de mesure; un dernier désavantage a trait à l'impossibilité de mesurer les dépendances séquentielles parmi les comportements puisque les données sont amassées de façon discontinue.

L'échantillonnage d'événements provient du codage continu dans le temps réel de tous les changements du comportement à observer. Chaque fréquence, ou durée, du comportement à coder durant la période d'observation est enregistrée. Les mesures de durée et de fréquence sont comparables lorsque les périodes d'observation sont de même longueur pour tous les sujets (ou pour le même) durant toute l'étude. Les méthodes de codage par échantillonnage d'événements s'associent avec un système d'enregistrement des données utilisant des moyens audio-visuels. C'est une façon de maximiser les chances d'observer tous les types de comportements qu'ils soient de durée momentanée ou significative, et qu'ils soient fréquents ou pas (Sackett, 1978). C'est donc une méthode plus précise, mais plus coûteuse.

Une troisième catégorie largement utilisée en recherche observationnelle, surtout dans les sciences exactes, est mentionnée dans Fassnacht (1982). C'est la méthode de codage basée sur des enregistrements de type dimensionnel; et par rapport à la recherche sur le comportement, elle a trait aux échelles d'évaluation des degrés d'intensité d'une réponse ou d'une même propriété chez différents sujets. En recherche sur le comportement humain, cette méthode fait appel à divers jugements subjectifs de la part de l'observateur et il est pratiquement impossible d'atteindre un bon niveau d'accord entre les observateurs sur les données.

Avec la classification de Altmann (1974), Beaugrand (1982) reprend les méthodes d'échantillonnage particulières et il décrit sept techniques d'échantillonnage qui peuvent être combinées ou agencées selon les questions que se pose le chercheur, ou selon les hypothèses qu'il formule.

La première que Altmann (1974) appelle échantillonnage «ad libitum», i.e. non structuré, est une observation informelle et non systématique. C'est l'observation préalable à toute observation systématique, où les objectifs du chercheur sont de repérer les événements plus prégnants et de découvrir des patrons naturels sans avoir pris de décisions conscientes d'échantillonner. Cette méthode ne permettant pas d'analyses quantitatives, il est rarement possible de déterminer la part de véracité ou de biais dans les différences observées entre les individus. Elle ne fait qu'identifier les régularités de base et suggérer des questions.

Une autre méthode d'observation informelle, dont les enregistrements portent sur les relations entre les individus, organise les données sur une matrice sociométrique. Le complètement de matrice a pour but de compiler le plus d'interactions dyadiques possible sans se conformer à un ordre d'observation ou à une durée uniforme. Pour chaque paire d'individus observés, on tient compte de la direction et du degré d'unidirectionnalité de la relation-cible. Des relations asymétriques de dominance, de proximité spatiale, de communication ou d'affiliation sont ainsi dégagées. Les cases de ces tableaux montrent des données de fréquences d'interactions dyadiques qui ne doivent cependant pas être comparées entre elles.

Une autre méthode plus détaillée et plus riche en information consiste à échantillonner les comportements d'un ou de plusieurs sujets, ou d'un groupe, de façon «continue et complète». Avec l'échantillonnage d'événements, les données recueillies par appareil (polygraphe ou éthographe) informent sur la nature du comportement, sur l'identité de l'acteur, sur le moment d'apparition et sur la durée des unités. La séance d'observation peut se dérouler sur le vif ou être enregistrée. Si cette dernière option est choisie, tous les autres types de codification peuvent être dérivés de l'enregistrement complet et continu. Aussi, deux ou plusieurs observateurs peuvent simultanément en faire la codification. Cette méthode permet d'effectuer une multitude de mesures comme les fréquences et les taux d'émission des comportements, les changements dans ces taux, les transitions dans les séquences,les transitions

et le synchronisme entre les individus et les interactions, les durées, les latences et l'importance relative de certains comportements.

L'échantillonnage par centrations successives se distingue de la technique précédente par la planification de focalisation sur un seul individu ou sur un groupe d'individus particulier durant une période limitée. Lorsque la durée d'observation prédéterminée est suffisamment longue et que le choix des patrons comportementaux est pertinent, cette technique procure des données relativement non biaisées pour une large variété de questions. Elle ne peut cependant répondre aux questions de synchronie comportementale et la restriction sur la quantité de catégories comportementales observables limite les généralisations.

L'échantillonnage en séquence est aussi une variation de l'échantillonnage continu et complet, mais avec le focus d'observation dirigé sur une séquence de comportements ou d'interactions, sauf qu'il ne peut être appliqué sur le vif. L'objectif est d'obtenir de l'information sur l'ordre ou la structure séquentielle des comportements ou des interactions. La notation s'exécute selon l'ordre d'apparition des événements jusqu'à l'interruption de la séquence. Les débuts et les fins de séquences sont déterminés par les faits mêmes plutôt que par des signaux externes. Ceci laisse apparaître une limite de cette méthode, puisque les choix de critères marquant le début et la fin d'une séquence peuvent manquer de constance et d'objectivité. Un autre biais est introduit par le fait de rechercher des événements organisés en séquence. De plus, les événements en longues séquences sont plus difficiles à identifier

et peuvent être ignorés ou perdus. Le seul intérêt est donc relatif à l'étude des transitions dans les séquences de comportement.

L'échantillonnage par présence ou absence, ou échantillonnage de temps, est une méthode systématique développée pour étudier le comportement spontané, surtout chez les enfants, durant de courtes périodes définies. Contrairement aux méthodes précédentes, c'est la notation simple de la présence ou de l'absence des états qui sera faite plus souvent que celle des événements. Durant une succession de courts intervalles réguliers déclenchés par un marqueur de temps, l'observateur note pour chaque intervalle si le comportement est apparu au moins une fois. Les séances d'une durée approximative de 20 intervalles sont habituellement répétées plusieurs fois durant la période d'étude et peuvent ainsi contribuer à fournir l'importance relative de certains comportements du répertoire à l'étude. Les avantages résident dans la facilité du codage et dans l'optimisation de l'accord des observateurs. Les inconvénients proviennent de la surinterprétation des scores, car ces derniers se limitent à des fréquences d'intervalles et non pas à des fréquences de comportements; de plus, le temps passé dans une activité ne peut pas être correctement estimé à partir d'un pourcentage d'intervalles.

L'échantillonnage par balayage instantané est un échantillonnage de temps à des moments pré-sélectionnés et théoriquement sans durée. L'observateur note simplement le comportement d'un sujet à un instant donné et il procède successivement si plusieurs individus sont à observer. Cette technique est utilisée pour estimer le pourcentage de temps que chaque

individu d'un groupe ou d'une espèce accorde à une ou plusieurs activités particulières. Elle est aussi utilisée pour obtenir une forme de recensement sur certains paramètres d'une large population. Les inconvénients découlent d'abord des difficultés à standardiser le temps d'observation pour chaque individu, ou phénomène, et à distinguer clairement les catégories à enregistrer. Ainsi, des biais d'attention relatifs aux sujets et aux activités privilégiées durant l'observation, et une durée d'observation non constante peuvent facilement s'imposer et donner des résultats faussés.

Les sections précédentes ont clairement illustré les préoccupations de la recherche observationnelle concernant la richesse des données et leur validité externe. En résumé, les choix que le chercheur exerce sur la taxonomie, le système d'enregistrement et la méthode de codage constituent des contrôles internes pour fournir et assurer des observations précises, valables et représentatives des régularités comportementales. Aussi, plusieurs facettes des données à recueillir sont des facteurs déterminants du degré de généralisation des résultats. Par exemple, la grandeur de l'échantillon (situations, cultures, moments de vie différents, individus) et l'uniformité des durées d'observation des sujets, de la situation d'observation et des périodes de la vie des sujets vont agrandir ou amoindrir la variabilité des observations. Le degré de validité externe des résultats en sera plus ou moins affecté. Aussi, l'objectivité et la fidélité des données obtenues par les méthodes d'observation directe font l'objet d'autres contrôles externes que nous commenterons au chapitre deux. Auparavant, pour mieux compléter la boucle dans la description

du processus observationnel, une description succinte et non exhaustive des principales méthodes d'analyse des données sera ébauchée.

D. Etape de l'analyse des données

L'analyse des résultats de l'observation, constituant la synthèse du processus observationnel, amène le chercheur à rendre compte de la véracité scientifique des phénomènes qu'il a observés et codés. Il a alors recours aux méthodes statistiques. Nous avons déjà parlé de certains apports de la statistique à l'observation comme dans les méthodes de cueillette et de gestion des données. Lors de ces étapes, la statistique est utilisée pour catégoriser le plus clairement possible les objets d'observation que l'on veut étudier par la suite. Par après, des statistiques descriptives sont utilisées pour condenser l'information; elles servent dans ce cas à compter des nombres d'occurrences correspondant à différentes classes d'observations. Bovet (1984) mentionne que la statistique répond alors aux besoins du dénombrement et du résumé de l'observation. Ces aspects rejoignent spécifiquement l'étape de l'analyse des données d'observation. Par rapport au dénombrement, Bovet (1984) mentionne les distributions d'effectifs et les distributions de fréquences; les histogrammes et diagrammes en bâtons sont les techniques privilégiées de représentation graphique. Lorsque plusieurs classifications du comportement ont été appliquées. l'ordinateur facilite le dénombrement.

Bovet ajoute que le résumé de l'observation conduit à la condensation des distributions mêmes et que parmi les procédures statistiques, on retrouve les

calculs d'indices. Aussi, chaque type de classification dans l'observation a son éventail spécifique d'indices. Bovet (1984) en décrit trois types (indices de tendance centrale, de dispersion et de liaison), que le mode de classification soit nominal, ordinal ou métrique.

Il existe plusieurs autres moyens statistiques pour résumer des observations. On peut se référer à Laurencelle (1986) qui en énumère toute une panoplie: construction de vecteurs de fréquences avec matrices transitionnelles, matrices de Markov pour enchaînement des comportements, séries chronologiques pour données dichotomiques, méthodes appliquées aux questions de dépendance séquentielle, de périodicité et de co-occurrence, méthodes d'estimations des taux et des durées des données échantillonnées à intervalles réguliers, analyse transitionnelle en décalages, représentation des probabilités conditionnelles progressives et régressives, etc. Bovet (1984) cite les techniques d'analyse factorielle, ou dimensionnelle, et d'analyse de correspondances en ce qui a trait à la caractérisation simultanée des liaisons entretenues entre des classifications. Sackett (1978) présente aussi divers types d'analyses pour mesurer le comportement pris dans son ensemble, les probabilités comportementales et les séquences. Il se base sur des méthodes d'échantillonnage systématique.

Sackett et Landesman-Dwyer (1982) identifient deux classes fondamentales d'analyses des données en recherche observationnelle: (1) les analyses non séquentielles qui impliquent le sommaire des données d'une séance d'observation (ou d'un ensemble de séances) complète; (2) les analyses

séquentielles qui ont trait aux relations temporelles entre les comportements d'un seul individu ou entre les comportements de plusieurs individus.

Beaugrand (1982) parle du calcul des probabilités de succession des comportements, i.e. l'analyse des transitions dans les séquences. Lui et ses collègues ont appliqué dans leur étude une analyse de la distribution des intervalles temporels et du logarithme de la survie des événements d'une même unité comportementale chez un même individu, ou de deux unités chez des individus différents.

Longabaugh (1980) fait ressortir que les problèmes d'analyse interne des données observationnelles proviennent de trois caractéristiques des études observationnelles:

(1) Le nombre des catégories comportementales est souvent très large; il en résulte un trop grand nombre de perspectives d'approche des données, ce qui accule l'investigateur au choix entre une analyse partielle en profondeur et une analyse superficielle de toutes les données. Longabaugh mentionne que les techniques de réduction des données les plus utilisées sont les analyses en regroupements hiérarchiques et les analyses factorielles. Elles impliquent divers procédés analytiques servant à examiner les interrelations parmi les catégories comportementales avant que chaque catégorie ne soit mise en relation avec des variables externes.

(2) La deuxième caractéristique concerne le fait que dans la plupart des études observationnelles, quelques-unes des catégories comportementales renferment des fréquences avec occurrences élevées, alors que la plupart des autres

catégories montrent des fréquences relativement faibles. Ceci entraîne la création d'une grande distorsion lors de la mise en graphique de la distribution des fréquences comportementales. Les catégories d'occurrences des comportements sont arrangées par ordre de fréquence sur l'axe horizontal, débutant avec la catégorie qui a le plus d'occurrences, et les fréquences réelles sont placées par échelle d'importance sur l'axe vertical. La courbe représentée montre alors une baisse rapide de la première à la deuxième catégorie avec, successivement, de plus petites baisses.

(3) La sensibilité ou la stabilité des données sont souvent rendues douteuses par la combinaison d'un faible nombre de sujets dans l'échantillon avec plusieurs comportements-cibles dans de multiples situations ou environnements. Il est donc difficile d'observer de vraies différences entre les sujets et il est encore plus difficile de procéder au résumé des données parce que la quantité de variables à l'étude dépasse largement le nombre de sujets.

Nous ajoutons enfin qu'une analyse des données comporte aussi une interprétation, ou synthèse, des résultats. Bovet (1984) explique que les interprétations des résultats de l'observation s'opèrent à partir de modèles statistiques faisant partie de l'ensemble de la statistique inductive, par opposition aux modèles de type descriptif servant à résumer l'observation. Dans un premier type d'inférence, la généralisation consistant à passer de l'échantillon à la population dont il est extrait s'effectue avec prudence, c'est-à-dire en ayant recours à des «limites de confiance». La statistique peut

conduire à un deuxième type d'inférence sur les données d'observation par l'usage des «tests statistiques» basés sur un calcul des probabilités.

Ce premier chapitre nous a donné un aperçu des stratégies observationnelles connues, nous faisant ainsi voir les différentes considérations dans la systématisation du processus d'enregistrement des données. En fait, nous avons vu que de multiples points sont à considérer pour obtenir des données fiables et le choix d'une méthode de codage conforme aux objectifs de la recherche permet au chercheur d'englober les aspects essentiels. Et malgré cette planification, la méthode de codage ainsi que le système d'enregistrement qui auront été élaborés avant la codification ne seront ni plus et ni moins valables que les codes qu'ils permettent d'emmagasiner par l'entremise de l'observateur ou du codeur. Ainsi, suite à l'évaluation des résultats d'observation, le chercheur peut reconsidérer son système d'enregistrement ainsi que ses méthodes d'échantillonnage. Cela signifie que le chercheur a intérêt à vérifier la valeur de ses données d'observation pour s'assurer de la pertinence de ses méthodes de cueillette de données. Cette vérification s'effectue au niveau des classifications résultant des interactions entre l'«instrument» d'observation (codeur) et son filtre (technique de codage).

Nous venons d'introduire le type de préoccupation constituant le fondement de la problématique de cette recherche et il sera plus longuement commenté au cours des deux prochains chapitres. Le chapitre qui suit traitera de la difficulté à rendre compatibles les concepts classiques de la théorie

psychométrique et leur opérationnalisation dans l'évaluation des données provenant des procédures d'observation directe.

Chapitre II

Les problèmes liés à la fiabilité des données en recherche observationnelle

Les problèmes méthodologiques liés à l'observation directe ont été commentés selon plusieurs points de vue (données d'observation, système d'observation, champ d'observation et temps d'observation) dans la section précédente. Cette section abordera le point de vue de l'observateur, plus particulièrement dans son interrelation avec les données d'observation.

Plusieurs auteurs (Johnson et Bolstad, 1973; Kazdin, 1977; Frame, 1979; Mitchell, 1979; Hollenbeck, 1980) ont passé en revue diverses sources d'influence sur le rendement de l'observateur, donc sur la valeur de ses observations. Les sources de biais couramment commentées dans la littérature sont ici énumérées sans plus d'élaboration[1]. Elles ont trait à la sensibilité de l'observateur dans la phase d'évaluation de sa performance, à la modification de performance d'un observateur selon qu'il est ou non en phase d'évaluation, à la complexité des comportements émis par les acteurs observés, à la qualité et à la quantité d'information transmise à l'observateur concernant sa tâche d'observation ou le contexte de recherche, aux attentes implicites de la part de l'observateur et de la part de l'investigateur, enfin aux feedbacks donnés aux observateurs.

---

[1] Déziel (1985) présente une discussion assez détaillée de ces aspects qu'elle a recensés à partir des mêmes auteurs ci-haut mentionnés.

Sackett et ses collaborateurs (1978) ramènent la non-fidélité de l'observateur à deux classes d'erreurs: les erreurs d'omission et les erreurs de commission. Pour ces auteurs, un système observationnel ne sera utile et opérationnel qu'une fois ces erreurs minimisées. Longabaugh (1980) situe trois étapes du processus observationnel nécessitant l'évaluation de la fidélité des codeurs: (1) au moment de vérifier le pouvoir des catégories à discriminer le comportement; (2) durant la période d'entraînement des codeurs, et (3) à l'étape de la véritable collection des données. Les études de fidélité se caractérisent selon les considérations particulières formulées à chaque étape. Néanmoins, le but poursuivi est le même pour toutes ces étapes: il s'agit de vérifier la pertinence des données obtenues relativement à l'objectif de décrire une «réalité», soit à travers le pouvoir de discrimination du code et du codeur, ou par la reproductibilité des données obtenues.

Les buts empirique et théorique découlant du rationnel dans la détermination du degré de fidélité d'une investigation particulière sont décrits par Sackett (1978); dans une perspective empirique, l'évaluation de la précision du système de codage est opérationnalisée en utilisant des statistiques qui mesurent le degré avec lequel la fidélité observée diffère du degré de fidélité maximale anticipé. Dans la perspective théorique, l'investigateur utilise des statistiques présentant un index du degré de non-fidélité ou de non-concordance présente dans les données dans le but de renseigner sur l'adéquation de l'instrument observationnel de codage. Ces statistiques présentent des estimés de la part de variance attribuable à l'imprécision de

l'observateur et à d'autres sources de non-fidélité, telle l'ambiguïté de certaines catégories pour un seul ou pour plusieurs observateurs. Ces estimés permettent de décider si le système de codage possède suffisamment de capacité discriminante pour réaliser sa fonction de mesure, même s'il existe toujours un certain degré d'infidélité.

Toujours dans le cadre de l'observation directe du comportement, Beaugrand (1982) désigne "fidélité instrumentale" l'indice de la fiabilité des observations produites par l'observateur. Faisant référence aux aspects affectant positivement ou négativement le rendement d'un observateur et donc la fiabilité de ses observations, il relève deux critères répondant de la pertinence des évaluations provenant d'un instrument: la constance (ou stabilité) et la justesse (ou précision) des mesures. Il mentionne qu'une étude de fidélité instrumentale doit d'abord être entreprise pour vérifier la validité de base d'un instrument de mesure, i.e. la précision et la stabilité des observateurs dans leurs réponses au même objet d'observation. Ses propos font ainsi le rapprochement entre la «fidélité instrumentale» et la conception psychométrique de l'erreur de mesure.

Dans le chapitre «mesure des phénomènes» de Bélanger (1982), il est mentionné qu'en plus de fournir des résultats pertinents, les instruments utiles doivent aussi procurer de l'information exacte et que l'exactitude de l'information scientifique repose sur l'objectivité, la fidélité, la validité et la sensibilité de la mesure. Bélanger explique comment les quatre facteurs d'exactitude de la mesure sont appréciés dans leur acception psychométrique

traditionnelle. Il n'apporte cependant pas de précisions sur les mesures résultant des procédures d'observation directe. Lorsqu'il aborde la question de la fidélité d'un instrument de mesure sous les trois aspects de stabilité, équivalence et homogénéité des mesures, il ne fait aucune spécification sur la particularité de l'instrument «observateur». De plus l'opérationnalisation de ces concepts relativement à l'évaluation de la performance des observateurs n'y est guère élaborée.

Or, comment dans le contexte de l'observation directe ces qualités de la mesure (dont Bélanger fait mention), i.e des résultats d'observation, sont-elles évaluées? Et qu'est-ce que ces évaluations métriques permettent de conclure sur les observations recueillies ou sur le procédé d'observation utilisé? Avant de discuter ces questions sous l'angle des multiples approches et méthodes d'évaluation des données, une première partie de nos propos abordera les théories psychométriques classiques telles qu'elles se définissent en théorie classique des tests et telles qu'elles sont appliquées actuellement dans le contexte d'observation directe. La deuxième partie, plus volumineuse, constituera une vaste revue des multiples utilisations de la fidélité en observation directe.

Première partie: Théories psychométriques appliquées aux données d'observation

Historiquement, le concept de fidélité a été rattaché à l'étude des différences individuelles et était restreint aux tests standardisés mesurant

des traits particuliers (intelligence, personnalité, etc.). Les intérêts de recherche sur le comportement humain évoluant vers des considérations développementales et fonctionnelles, les méthodes d'observation se sont aussi transformées pour se situer de plus en plus dans un contexte naturel ou quasi-naturel. Avec les méthodes d'observation directe, l'enregistrement du comportement est effectué par des observateurs humains, et ces derniers sont devenus les nouveaux instruments dont on doit évaluer la fidélité; les concepts classiques de la fidélité sont encore utilisés avec les mesures d'observation directe (Johnson et Bolstad, 1973; Mitchell, 1979). L'instrument de mesure qui était le test ou le questionnaire, mesurant indirectement des traits ou des caractéristiques, est devenu l'observateur humain, mesurant directement un ou plusieurs comportements. Les quelques aspects métriques de la recherche observationnelle sur le comportement humain qui viennent d'être introduits nécessitent donc une explicitation de leurs origines et de leurs définitions théoriques, lesquelles seront apportées dans le volet qui suit. Le deuxième volet de cette section traitera des divers points de vue relatifs à l'application des concepts classiques à l'observation directe.

Définition des concepts de la théorie des tests

Dans la théorie classique des tests, la fidélité d'un instrument est évaluée à partir de l'erreur de mesure qu'il produit, ou de la consistance et de la stabilité des scores des individus relatifs au trait, à la caractéristique ou au comportement mesuré (Mitchell, 1979).

Bernier (1985) explique que la théorie classique des tests apporte certains éléments de réponse aux problèmes concernant le degré de consistance des résultats et les causes des différences entre les scores. Tout score «observé», le résultat à un test, se compose additivement d'un «score vrai» et d'une «erreur». L'auteur rappelle qu'il y a plusieurs façons d'illustrer la relation entre le construit hypothétique «score vrai», et le score observé. Ainsi, une façon de concevoir le «score vrai» est de le représenter comme le résultat à un test parfait, i.e. sans erreur. Une autre façon le décrit comme le résultat moyen d'un individu sur un même test administré un nombre infini de fois. Le concept de l'erreur de mesure, pour sa part, a trait à la surestimation ou à la sous-estimation du score vrai par le score observé, l'erreur étant indifféremment positive ou négative. L'erreur de mesure est donc définie comme la différence entre le score vrai et le score observé, i.e. une différence essentiellement due ou attribuée au hasard.

Deux modèles permettent d'évaluer l'importance de l'erreur de mesure: celui de la corrélation entre des tests parallèles et celui de la proportion de variances provenant de l'échantillonnage d'un univers. Le premier est plus populaire parce qu'il suffit d'obtenir la corrélation entre deux tests parallèles pour déterminer la fidélité d'un test. Le principe formulé est que les erreurs aléatoires dans un test sont en corrélation nulle avec les erreurs d'un autre test. Après une démonstration mathématique de réconciliation entre les deux modèles, Bernier conclut que "le coefficient de fidélité correspond à un rapport de variances et prend des valeurs entre (0) et (1). (...) La fidélité est la

certitude que l'on a qu'un test mesure des scores vrais et le coefficient de fiabilité nous en donne le degré" (p. 100).

(F1)

$$\text{Coefficient de fidélité} = \frac{\text{variance des scores vrais}}{\text{variance des scores observés}}$$

ou

$$= 1 - \frac{\text{variance des erreurs}}{\text{variance des scores observés}}$$

Par contre, la corrélation entres les scores vrais et les scores observés (i.e. entre des mesures qui donnent précisément le même score vrai à chaque individu) nous donne un «indice de fidélité», lequel est égal à la racine carrée du coefficient de fidélité. Cet indice spécifie la précision avec laquelle les items permettent de mesurer quelque chose.

(F2)

$$\text{Coefficient de corrélation}^2 \text{ entre scores vrais et scores observés} = \text{Coefficient de fidélité}$$

Pour estimer le coefficient de fidélité, ou la précision des mesures, la théorie classique des tests fournit des façons de calculer l'erreur de mesure due au hasard, tout en postulant que l'erreur aléatoire n'est jamais

complètement éliminée de toute mesure. Le postulat hypothétique implique "la possibilité d'obtenir les mêmes résultats pour les mêmes personnes dans les mêmes conditions et à des occasions différentes, mais avec des tests mesurant le même trait", i.e. "la consistance d'un ensemble de mesures" (Bernier, 1985; p. 102). L'idéal serait de faire un grand nombre de mesures sur un seul individu mais en pratique on doit tester plusieurs individus deux fois; la différence entre leurs rendements est utilisée pour estimer la fidélité d'un test.

Les trois catégories d'erreurs identifiées par Bernier (1985) sont: (1) les erreurs dues au test lui-même, (2) les erreurs reliées à l'administration du test et, (3) les erreurs reliées aux répondants du test. Les deux premières catégories représentent des sources d'erreurs qui sont facilement contrôlées en limitant l'influence instabilisante de plusieurs facteurs et en standardisant l'administration des tests; il est par contre plus difficile d'éviter les sources d'erreur de la dernière catégorie puisqu'elles sont de type intra-subjectif. Bernier (1985) énumère quelques-uns de ces facteurs faussant les résultats des répondants: "la motivation, l'habitude des tests, l'anxiété, l'apprentissage différentiel et les variables d'ordre physiologique" (p.104).

Ainsi, s'il y a plusieurs sources d'erreurs possibles, il y aura plusieurs méthodes d'estimation de la fidélité. Trois méthodes d'estimation sont identifiées comme suit par Bernier (1985):

(1) Méthode des tests parallèles, laquelle consiste à estimer la corrélation entre deux versions d'un même test composé d'items différents et administrées

à un même groupe avec un intervalle de temps minimum, de façon à réduire les fluctuations possibles entre les scores vrais aux deux passations. Le manque de parallélisme entre les deux versions sera traité comme de l'erreur et contribuera à réduire le coefficient de corrélation. Le coefficient de fidélité obtenu par cette méthode peut servir de mesure du degré d'équivalence entre les deux versions.

(2) Méthode du test-retest, laquelle consiste à administrer un même test deux fois successives au même groupe d'individus. Cette méthode comporte un problème d'effets combinés entre l'intervalle de temps (d'une passation à l'autre) et la mémoire. Si l'intervalle entre les deux passations est très court, l'effet de mémoire peut influencer les répondants et mener à une forte corrélation entre les résultats, entraînant la surestimation de la fidélité. Par contre, si l'intervalle est trop long, les modifications historiques dans l'habileté des répondants, i. e. les scores vrais, vont avoir pour effet de réduire la corrélation, entraînant la sous-estimation de la fidélité (Bernier, 1985: p. 112). Par cette méthode, on obtient des coefficients de stabilité qui permettent d'apprécier combien la caractéristique mesurée est stable dans le temps, i.e. s'il n'y a pas d'effets de pratique et d'apprentissage différentiel (p. 118).

(3) Méthode de l'analyse interne de la variance: elle vient pallier aux difficultés des deux méthodes précédentes. C'est la méthode des moitiés, où l'on procède à construire deux moitiés d'un même test en équilibrant le degré de difficulté de chacune après avoir placé les items par ordre de difficulté. Les

items de chaque moitié peuvent soit être répartis au hasard, soit être pairés selon leurs indices de difficulté, de façon à construire deux tests de même difficulté moyenne et de même variabilité. Ensuite, on calcule la corrélation des scores des individus à chaque moitié du test. Cette estimation de la fidélité d'une moitié de test doit par la suite être corrigée pour s'appliquer au test en entier, en utilisant une formule appropriée dite de Spearman-Brown. Cette méthode a l'inconvénient de surestimer quelque peu la fidélité du test, parce qu'elle mise sur un contexte simplifié de l'application du test. Le coefficient obtenu par cette méthode s'interprète de la même manière qu'un coefficient d'équivalence. On peut aussi obtenir un coefficient d'homogénéité, ou de consistance interne, en calculant le coefficient «Alpha» de Cronbach ou en appliquant la «Formule 20» de Kuder et Richardson. Chaque item du test a un coefficient de fidélité comme si chacun avait été échantillonné au hasard et on considère que l'on obtient une corrélation entre des items (tests) parallèles.

Bernier (1985) conclut sur ces méthodes en mentionnant que la combinaison des méthodes «test-retest» et «tests parallèles», c'est-à-dire en administrant deux versions de test au même groupe d'individus et en laissant un intervalle entre les deux passations, procure une estimation sévère de la fidélité; le coefficient de stabilité-équivalence obtenu constitue un seuil inférieur de la fidélité ou de la précision du test. Aussi, il ajoute qu'il y a plusieurs coefficients de fidélité possibles pour un même test signifiant différentes choses; alors, on ne peut parler de «*la fidélité*» d'un test.

D'autres facteurs influencent le coefficient de précision d'un test et Bernier énumère les suivants: (1) l'étendue des différences individuelles (un groupe homogène donne une sous-estimation de la fidélité comparativement à un groupe varié ou hétérogène); (2) la difficulté du test (le degré de difficulté d'un test devrait être tel qu'on obtienne une grande distribution de scores pour ainsi maximiser la précision); (3) la longueur du test (plus d'items entraînent plus de précision); et (4) la limite de temps (si une vitesse est imposée, la méthode des moitiés est inefficace).

La question qui fait suite à l'estimation de la fidélité est liée à l'interprétation du coefficient obtenu. Le coefficient de fidélité calculé est-il acceptable? Bernier mentionne quatre façons d'interpréter un coefficient de fidélité:

(1) Comme corrélation entre scores obtenus et scores vrais : ici, on interprète la fidélité comme la proportion de variance totale qui est de la variance vraie (e.g.: si r_{XX} = .50, on peut dire qu'il y a 50% de variance vraie dans la variance des scores observés); cette méthode permet d'apprécier "le degré d'erreur de mesure obtenu en administrant une forme d'un test à un échantillon particulier d'individus dans certaines conditions bien spécifiques" (Bernier, p.135).

(2) En comparant la précision d'un test avec la précision d'autres tests de même type: on compare un coefficient de fidélité d'un test avec d'autres coefficients de tests semblables pour obtenir un standard de comparaison (e.g.: les tests d'aptitude atteignent habituellement un seuil de 0,90 en fidélité, alors que pour les tests de personnalité on observe des corrélations autour de 0,80).

(3) Comme le pourcentage de personnes qui changeraient de rangs: en se référant à Thorndike et Hagen (1961), Bernier illustre le changement de rangs d'un individu en fonction des diverses valeurs que peut prendre la fidélité à différentes administrations d'un test; ainsi, plus la précision est parfaite, moins il y a de chances qu'un individu change de rang aux différentes administrations du test; l'inverse se produit quand r_{XX} prend une faible valeur.

(4) Comme un indice de la quantité d'erreur dans les scores individuels : l'indice de dispersion des scores observés autour du score vrai est appelé «erreur-type de mesure», $SE = SX \sqrt{1 - r_{XX}}$, SE étant l'erreur-type de la mesure, SX l'écart-type des mesures ou scores observés, et r_{XX} le coefficient de fidélité; cette valeur détermine l'intervalle moyen dans lequel peut se retrouver le score vrai. Ainsi, après avoir évalué le degré de variabilité des scores observés d'un individu autour de son score vrai, on se sert de la courbe normale de probabilité pour interpréter.

La validité est une autre propriété importante de la mesure et elle est en interrelation avec la fidélité d'une même mesure, i.e. elle est toujours plafonnée par la fidélité du test. Toutefois, une fidélité élevée n'est pas garante d'une forte validité, puisque la validité est toujours spécifique à une situation donnée. La validité concerne ce qui est mesuré ainsi que la qualité de la mesure: «Qu'est-ce que le test mesure?» et «A quel degré le test mesure-t-il bien?». Dans la théorie des tests, on définit la validité comme étant la "portion de variance vraie qui est pertinente avec les buts de l'utilisation d'un test" (Bernier, p. 171). Cependant, une faible variance d'erreur (ou variance des erreurs) n'assure pas une forte validité, car tout dépend de la pertinence de la variance vraie. La définition de la validité s'illustre ainsi:

(F3)

Variance vraie = Variance pertinente + Variance non-pertinente

et la variance totale d'un test sera composée de la variance vraie décomposée en facteurs et de la variance d'erreur:

(F4)

Variance totale	=	Variance pertinente	+	Variance non-pertinente	+	Variance d'erreur

L'étude de la validité comporte deux approches: l'une qui "se préoccupe de la qualité avec laquelle un test mesure un trait hypothétique ou un construit", l'autre qui s'intéresse à "la relation entre les scores obtenus à un

test et une mesure-critère" (Bernier, p.171). En théorie des tests, on parle de trois types de validité: validité critérielle, validité de contenu et validité théorique (de construit). La validité critérielle utilise le score au test comme prédicteur d'une performance future ou d'une position sur un critère en mesurant le degré de relation ou d'association entre les scores au test et au critère (Ex.: résultats scolaires et aptitudes pour une profession spécifique). Johnson et Bolstad (1973) expliquent que la validité critérielle est établie quand la mesure d'une dimension comportementale est en corrélation avec un critère établi par un instrument de mesure différent. Dans ce type de validité, le critère est plus important que le contenu du test. La validité de contenu consiste à faire un jugement sur le degré de pertinence de l'échantillon d'items du test. Ici, l'importance est mise sur le contenu du test car on postule que celui-ci est un échantillon représentatif de l'univers des situations circonscrites. La validité théorique ou «de construit»s'intéresse à connaître le trait que le test mesure; on veut savoir si le test mesure bien les traits hypothétiques ou les qualités présumément réfléchies dans la performance au test. L'estimation de cette validité s'opérationnalise par l'accumulation des preuves sur le trait mesuré; elles peuvent provenir de plusieurs sources, dont les études de validité critérielle et de contenu. Johnson et Bolstad (1973) parlent de validité «convergente», ou «concurrente» (Bernier, p.179): il s'agit d'une validité critérielle qui est établie lorsque deux méthodes de mesure différentes de la même variable donnent des résultats similaires ou en corrélation.

Un résumé des concepts de fidélité et de validité ci-haut expliqués est fourni par Runkel et McGrath (1972). Le concept de validité concerne l'aspect qualitatif du modèle se trouvant entre le concept et l'opération. Ainsi, la question de validité fait appel à la probabilité que la mesure utilisée dans la définition opérationnelle soit «vraiment» une mesure de la propriété correspondante telle que conceptuellement définie. La question de fidélité pose le problème de savoir si la mesure utilisée en tant que définition opérationnelle constitue une garantie de produire la même valeur dans des évaluations indépendantes répétées du même «objet». Runkel et McGrath déclarent que toute mesure implique l'introduction d'une certaine part d'erreur, qu'elle soit aléatoire ou systématique. Il en résulte la nécessité d'évaluer la grandeur de cette erreur pour apprécier la précision de la mesure. Le dilemme rencontré est d'obtenir à la fois la reproduction du résultat et l'applicabilité de ce résultat dans un éventail de conditions. Runkel et McGrath rappellent que la notion de généralisabilité contenue dans un aspect de la fidélité (applicabilité) a été élargie par Cronbach, Rajaratnam et Glaser (1963; dans Runkel et McGrath, 1972).

La théorie de la généralisabilité développée par Cronbach et ses collaborateurs (1972; voir Berk, 1979) emprunte ses méthodes statistiques des procédures d'analyse de la variance dues à Fisher. L'étude de généralisabilité sert à apprécier diverses sources de variation (ou facettes[1]) dans les scores.

---

[1] Un «univers» est obtenu en combinant des facettes; c'est à partir des univers de généralisation que le chercheur explore les sources d'erreurs. On peut se référer à Déziel (1985) pour un résumé des

Les coefficients de généralisabilité ainsi obtenus fournissent des estimations du degré auquel les scores observés sont confondus avec l'erreur. Ils indiquent la généralisabilité d'un échantillon d'observations à un univers spécifique d'observations. En d'autres mots, la formulation mathématique de l'index de généralisabilité permet de connaître les limites de généralisabilité des résultats d'un ensemble d'observations, c'est-à-dire le degré de similitude et de différence des résultats en considérant un univers d'observations potentielles défini de façon générale et hétérogène (selon les formes, les conditions et les occasions de mesure). Par contre, en posant les questions reliées aux concepts de fidélité et de validité, Runkel et McGrath présument que l'on restreint les informations sur la mesure aux deux aspects respectifs suivants: (1) le niveau de ressemblance entre deux ensembles d'observations presque identiques et avec les mêmes formes et conditions de mesure ainsi que la même population de répondants; et (2) le niveau de ressemblance d'un ensemble d'observations tirées d'un univers considéré comme un ensemble-critère d'observations potentielles. Runkel et McGrath en concluent que la standardisation d'un contexte est inversement reliée à sa généralisabilité: lorsqu'on tente d'améliorer la précision (dans le sens de répétition) d'un échantillon d'observations en réduisant les variations dans cet échantillon, on contribue à diminuer sa généralisabilité puisque le nombre de sous-échantillons a été restreint; de même, lorsqu'on veut agrandir l'échantillon tiré d'un «univers» pour augmenter sa généralisabilité, la variabilité dans les sous-

---

principaux univers de généralisation concernant les études d'observation. Mitchell (1979) donne des explications plus détaillées.

échantillons s'en trouve augmentée et ceci a pour effet d'obscurcir la précision de ce que l'on veut détecter.

La section suivante nous présente comment les chercheurs oeuvrant dans le contexte de l'observation directe tentent de reprendre les concepts classiques de fidélité et de validité pour apporter un fondement scientifique à leurs méthodes empiriques. Nous pourrons constater que l'unanimité est loin d'être atteinte relativement à la transposition des concepts de la théorie classique des tests; certains n'hésitent pas à reconnaître que les mesures provenant des tests sont différentes de celles de l'observation directe, présage que l'applicabilité théorique de ces concepts est problématique et leur opérationnalisation pas moins difficile.

Application des concepts classiques à l'observation directe

L'observation directe est la base de l'analyse du comportement ou de l'évaluation comportementale. Les chercheurs ont initialement accepté cette «méthode» comme valide par définition: Kent et Foster (1977) mentionnent que l'enregistrement de l'observation du comportement fut considéré comme le «portrait scientifique» le plus pur, vision qui a prévalu jusqu'aux années 1970. Depuis une quinzaine d'années, la qualité des données produites par les procédures d'observation directe a reçu de plus en plus d'attention (Cone, 1977; Johnson et Bolstad, 1973; Kent et Foster, 1977). Ces remises en question sont suscitées par un argument fondamental stipulant que les découvertes de la

recherche ne peuvent être plus valides ou fidèles que les méthodes de mesure sur lesquelles elles se basent. La littérature béhaviorale touche particulièrement à la question de l'accord inter-juges, où chacun tente d'apporter sa lumière sur les controverses concernant la façon de calculer les indices d'accord pour établir la pertinence et la qualité des observations obtenues. On y retrouve une vaste documentation concernant l'effet des variations expérimentales sur les taux d'accords inter-juges et sur la fiabilité des données d'observation. Les nombreuses études portant sur les biais des observateurs, les attentes, la complexité du code et du système de codage, la sensibilité de l'observateur à l'évaluation de sa performance, etc. sont reliées à ces questions (voir Kazdin, 1977a; Wildman et Erikson, 1977; Hartmann, 1977; Johnson et Bolstad, 1973; Kent et Foster, 1977; Beaugrand, 1982).

L'usage des techniques observationnelles s'est généralisé à plusieurs situations tout en devenant de plus en plus spécifique; ceci renforce le besoin de rendre compte de la valeur de la méthodologie observationnelle utilisée. Et si l'«art d'observer» s'est développé avec de plus en plus de sophistication, les méthodes d'évaluation des données n'ont pas reçu le même degré d'attention. Les aspects méthodologiques des systèmes observationnels ayant été moins examinés incluent la précision des données observationnelles et les préoccupations de validité. En 1978, Hollenbeck déclarait que peu de systèmes implantés rencontraient les standards de mesure tels qu'appliqués à d'autres mesures psychologiques utilisant les théories classiques de fidélité et de

validité; pour lui, la rigueur dans l'usage des concepts de fidélité devient impérative pour les chercheurs employant les méthodes observationnelles.

Dans la plupart des recherches appliquées, les expérimentateurs tentent d'augmenter la probabilité que les données reflètent fidèlement le comportement du sujet observé en évaluant le degré auquel deux observateurs s'accordent sur le fait qu'une réponse comportementale spécifique s'est produite. D'autres auteurs rapportent cette comparaison comme un indice de la fidélité d'observation. L'étape consistant à déterminer la précision des procédures d'observation directe doit s'effectuer avant que les données ne soient utilisées pour produire l'information sur le comportement. Kent et Foster (1977) expliquent que de fréquents estimés de la concordance des jugements simultanés des paires d'observateurs sont nécessaires pour affirmer que les enregistrements comportementaux sont le produit reproductible des procédures d'enregistrement bien définies plutôt que les jugements idiosyncratiques de plusieurs observateurs. Ces échantillons d'accords entre observateurs, ou de corrélations entre observations, sont souvent identifiés comme la «fidélité» puisqu'ils sont destinés à refléter la qualité des enregistrements comportementaux obtenus au cours d'une investigation particulière. Peu importe les termes utilisés, la plupart des auteurs qui rapportent l'accord plutôt que la corrélation, utilisent des variations de la même procédure. En général, chaque session d'observation est répartie par unités de temps fixes; le nombre d'unités de temps marquant un accord entre

deux observateurs est divisé par la somme des accords et des désaccords, puis ce quotient est multiplié par 100. Le résultat est rapporté en tant que pourcentage d'accords entre les observateurs. Quoique bien des variations existent à l'intérieur de cette procédure générale, les plus communes sont les variations dans la longueur des unités de temps et dans la définition d'un intervalle d'accord. Plusieurs de ces variations seront décrites à la section suivante. Baer (1977) résume cette application de la fidélité aux données d'observation comme suit: il définit le pourcentage d'accords des mesures d'intervalles comme une représentation du nombre de fois que deux observateurs équipés avec les mêmes définitions du comportement voient ce comportement apparaître, ou ne pas se produire, durant les mêmes intervalles d'observation d'un sujet. La «fidélité» inter-observateurs dans cette vision, donne un estimé de l'erreur due aux observateurs seulement; cependant, les différences entre observateurs ne sont pas les seules sources d'erreurs des études observationnelles, et plusieurs auteurs (Laurencelle, 1986; Berk, 1979; Mitchell, 1979; Hollenbeck, 1978; Sackett et al., 1978; Hartmann, 1977; Johnson et Bolstad, 1973) insistent pour que la précision des données d'observation soit analysée sous d'autres aspects (variation dans les sujets ou objets observés, variation dans les circonstances, etc.) et en tenant compte d'autres facteurs (part d'accord imputable au hasard, probabilité, fréquence et transition des comportements, etc.).

En 1973, Johnson et Bolstad faisaient déjà la distinction entre «consensus d'observation» et «précision d'observation», le premier se basant

sur le calcul de la concordance des observateurs et la deuxième étant évaluée en comparant le résultat de codage d'un observateur avec un critère de codage établi au préalable. De plus, ils mentionnaient que, dans le sens traditionnel de la fidélité, l'examen de ces deux aspects n'était pas suffisant pour s'assurer de la fidélité d'une mesure. La notion de fidélité classique implique en plus une consistance de la mesure dans le temps (fidélité test-retest) ou sur des échantillons d'observation obtenus au même moment (fidélité des moitiés). Aussi, nous verrons dans la prochaine section comment l'évaluation de la fidélité des observateurs peut être faussée par les paramètres qui la définissent; l'estimé de fidélité obtenu sera parfois rehaussé, atténué, ou simplement farfelu.

La notion de fidélité n'a pas atteint une définition théorique faisant l'unanimité parmi les chercheurs. Berk (1979) croit que le terme «fidélité» est une fausse appellation; il se base sur la définition psychométrique classique, i.e. le rapport des composantes de variance du score observé et du score vrai, pour constater qu'elle n'est pas transportée dans la mesure de fidélité inter-juges. A son avis, le terme «accord» est plus approprié et représente mieux les statistiques figurant dans la littérature; le pourcentage d'accords, lorsque pris pour indice de fidélité, reflète l'efficacité de l'entraînement de l'observateur et le degré d'objectivité avec lequel le comportement-cible peut être mesuré. Par contre, Berk affirme que la reproductibilité des données entre observateurs n'est pas une assurance de leur validité; la cohérence dans l'observation peut aller de pair avec de l'inexactitude et il est donc nécessaire d'évaluer cette

autre propriété, essentielle à la qualité d'un système observationnel. Cependant, nous référerons aux travaux de Cone (1982) pour commenter l'aspect de la validité des mesures observationnelles.

Cone (1982) considère que le simple fait de déclarer une mesure valide parce qu'elle mesure ce qu'elle est censée mesurer constitue une forme limitée de validité. En observation directe, cette vision a longtemps prévalu comme une propriété acquise des mesures ainsi recueillies; ceci a eu pour effet qu'on a négligé l'étude des autres aspects de la validité. Cone insiste sur l'insuffisance de ce critère de base de la validité, lequel il préfère nommer «justesse», pour se prononcer sur l'adéquation des mesures observationnelles; à son avis, la justesse d'une mesure n'assure pas sa validité et la preuve d'une relation entre une mesure comportementale avec d'autres mesures du même comportement est nécessaire:

> Un système d'observation directe qui mesure ce qu'il est supposé mesurer, i.e. qui reflète fidèlement les caractéristiques objectives et topographiques du comportement, est dit être «juste». Si les scores sur un tel système sont reliés aux scores sur les mesures d'autres caractéristiques, le système est dit être «valide». Les systèmes utilisés de façon cohérente sont considérés fidèles. Un système ne peut être juste et non-fidèle, mais les systèmes peuvent être valides et non-justes, et ils peuvent être fidèles et non-justes (p.69).

Cone (1982) en conclut que l'étude incomplète de la validité dans les données d'observation directe est due à la richesse et à l'ambiguïté de ce qu'on

mesure par l'observation directe. Selon lui, l'incertitude existant au niveau de l'applicabilité des notions traditionnelles de validité provient d'une traduction inadéquate de celles-ci dans le contexte de l'observation systématique.

Les propos de Hollenbeck (1978) viennent appuyer ceux de Berk et de Cone: Hollenbeck précise que la fidélité observationelle doit, avec le contrôle des conditions d'observation, englober aussi la «justesse» et la «stabilité» des mesures. Il réfère aux définitions de Kerlinger (1964) pour décrire ces deux composantes de la mesure: si la mesure est une vraie représentation de ce qui est observé, elle est dite «juste»; si quelqu'un obtient le même résultat sur des mesures répétées utilisant des instruments similaires dans les mêmes conditions, le critère de «stabilité» est rencontré. Pour Hollenbeck, la mesure de l'accord inter-observateurs n'a de valeur au plan fidélité qu'une fois qu'elle a été comparée à un standard préétabli et aussi qu'elle a été calculée sur plusieurs essais répétés. La condition «justesse» implique l'utilisation d'un «codeur-expert», dont les observations serviront de «critère de vérité» aux observations subséquentes; la condition «stabilité» sera démontrée par le consensus des observateurs sur de multiples essais.

Les discussions qui précèdent révèlent le caractère d'interdépendance des notions de fidélité et de validité et nos propos se poursuivent avec quelques autres points de vue, relativement à l'applicabilité de ces concepts. Cone (1977) affirmait que les concepts de fidélité et de validité utilisés dans l'évaluation traditionnelle étaient aussi bien applicables à l'évaluation béhaviorale. Les différences entre les deux approches étaient perçues par Cone

comme étant plus philosophiques que méthodologiques. Il expliquait cette divergence par le fait que l'investigateur traditionnel met le focus sur les comportements par des méthodes d'observation indirecte et, en contrepartie, que l'investigateur béhavioriste centre directement son attention sur les réponses observables, i. e. sur des échantillons de ces réponses repérés dans l'environnement naturel. Ainsi, dans le premier cas, les différences comportementales observées sont expliquées en termes de traits hypothétiques; dans le deuxième cas, on se défend d'utiliser des construits hypothétiques en choisissant des procédures qui mesurent les comportements évidents de façon continue, avec des unités standard et absolues; les différences dans le comportement sont alors considérées comme l'ouverture qui nous permet d'observer les variables contrôlant ce comportement.

Baer (1977) apporte un autre point de vue de l'analyse de la fidélité en faisant ressortir le contraste entre l'évaluation des instruments psychométriques et celle des systèmes observationnels. Dans le premier cas, le questionnaire représente le modèle du problème de l'analyse de fidélité et la préoccupation est de savoir si l'instrument en soi mesure vraiment quelque chose; les items (questions) sont construits pour apparaître homogènes dans la même dimension de mesure, et l'essence du score de fidélité sera donc l'homogénéité des items combinés pour produire ce score. Avec les systèmes observationnels, les intervalles échantillonnés constituent les items et il n'y a pas d'obligation d'homogénéité entre les intervalles puisque l'observateur tente de répondre à la même question ouverte: «Qu'est-ce que le sujet fait

maintenant?». L'observateur utilise sa définition du comportement à chaque intervalle et ce comportement n'est pas distribué de façon homogène parmi les intervalles; ainsi, les «items» ne peuvent pas être construits mais plutôt acceptés comme ils se présentent tout en ayant la possibilité d'évoquer quoi que ce soit. La conclusion de Baer est la suivante: l'homogénéité est requise parmi les observateurs et les mesures de l'accord entre observateurs sont pertinentes au problème de fidélité des systèmes observationnels; le nombre d'observateurs est l'aspect à améliorer. Toutefois, dans cette vision pratique, Baer ne prend pas en considération les situations où la mesure d'accord n'atteint pas un niveau élevé; même s'il n'y a pas d'homogénéité évidente sur le choix des catégories observées, la mesure d'accord demeure pertinente et le problème est alors porté à l'aspect conceptuel de la fidélité. «Peut-on envisager que la conception classique de la fidélité soit revisée pour considérer la concordance entre observateurs comme une vérification appropriée au contexte de l'observation directe?»

Hartmann (1977) prétend que le principal intérêt de plusieurs chercheurs se résume à la fidélité de leur système de base d'acquisition des données, soit l'observateur humain; cette fidélité est définie comme le degré auquel un ensemble donné d'observations peut être généralisé aux observations que d'autres codeurs peuvent produire. Il souligne que la fidélité des données observationnelles peut être examinée à partir de nombreuses perspectives, telles que la cohérence des intervalles et la stabilité dans le temps, à travers les situations et le comportement. Par contre, Laurencelle (1983, 1986)

explique que deux aspects de fidélité sont impliqués dans le processus observationnel: on se questionne sur la fiabilité des observateurs et sur la fiabilité des données. Pour cet auteur, deux types de mesure sont donc requis pour apprécier ces deux aspects: dans le premier cas, la mesure sert à déterminer la «vraie compétence moyenne» de l'observateur et dans l'autre, on évalue le «contenu de vérité» des données. Ce dernier aspect est lié de près à la question de validité.

Les propos de Mitchell (1979) nous fourniront la synthèse de cette section. Cet auteur identifie trois façons de concevoir la fidélité des données observationnelles: (1) le degré auquel deux observateurs travaillant indépendamment s'accordent sur les comportements qui se sont produits; dans ce cas, les chercheurs rapportent le coefficient reflétant ce degré d'accord; le coefficient de l'accord entre observateurs reflète l'objectivité des différents observateurs utilisant la même méthode d'enregistrement du même comportement, mais sa détermination ne suffit pas à fournir toute l'information sur la qualité des données observationnelles; (2) la mesure observationnelle peut être considérée un cas spécial des tests psychologiques standardisés et les méthodes d'estimation de fidélité de la psychométrie classique sont utilisées (test-retest, formes alternées, etc.); ces coefficients de fidélité renseignent sur la stabilité et la consistance des différences individuelles parmi les sujets; ils confondent l'erreur de mesure avec d'autres sources de variabilité; et (3) la mesure observationnelle peut être pensée comme fournissant des données qui sont sous l'influence de nombreux aspects

différents dans la situation d'observation (ex.: différents observateurs ou différentes occasions), incluant les différences individuelles entre sujets; ce troisième point de vue a trait à la théorie de la généralisabilité; le coefficient de généralisabilité procure aussi de l'information sur la stabilité et la consistance des différences individuelles, mais il est supérieur par son aptitude à tenir compte des sources de variance autres que les différences individuelles et que l'erreur de mesure.

En conclusion, nous pouvons observer que l'utilité des concepts fondamentaux de la théorie classique devient controversée quand on tente de les opérationnaliser pour les mesures d'observation directe. Campbell et Fiske (1959; voir Johnson et Bolstad, 1973) définissent leur opérationnalisation comme suit:

> La <u>fidélité</u> est la concordance entre deux efforts à mesurer le même trait par des méthodes similaires au maximum. La <u>validité</u> est représentée dans l'accord entre deux tentatives de mesurer le même trait par des méthodes différentes au maximum (p.50).

Les différentes propositions pour rendre les concepts classiques opérationnalisables ont chacune leurs lacunes pour une raison première, dont un des auteurs fait mention (cf. Baer, 1977): le contexte de l'observation directe est différent sur plusieurs aspects, et ces distinctions ne sont pas considérées comme telles dans l'application des théories classiques à ce domaine de recherche. Ces propos nous introduisent au coeur des problèmes soulevés par la

vérification des méthodes d'observation directe, i.e. l'utilisation de concepts théoriques appropriés aux procédures d'analyse de la qualité des données d'observation directe. La prochaine partie présentera une multitude de procédés d'évaluation des données, mais elle introduira au préalable des définitions visant à situer plus clairement ce domaine jusqu'à présent assez épars.

Deuxième partie: Multiplicité des approches de la fidélité des données d'observation directe

La littérature de recherche dans le domaine de l'observation directe du comportement humain fait état de l'inconfort des chercheurs à choisir une méthode statistique de vérification des données. Il n'existe pas à ce jour une conception théorique de la fidélité intégrée aux données observationnelles et faisant consensus. L'absence de méthodes standardisées pour calculer l'accord inter-juges est un effet de cet état de choses. Plusieurs chercheurs ont proposé différentes solutions; Harris et Lahey (1978) relatent des études montrant comment les publications de la fin des années 1970 étaient orientées sur ce problème: Baer, 1977; Kelly, 1977; Hartmann, 1977; Hopkins et Hermann, 1977; Kratochwill et Wetzel, 1977; Yelton, Wildman et Erikson, 1977. L'objectif poursuivi ici sera donc de fournir un aperçu de la fécondité des propositions et de montrer leurs avantages et limites. Le sommaire présenté ci-après a été organisé à la suite d'une vaste revue de littérature qui avait pour but premier de trouver des études de fidélité portant plus spécifiquement sur des données enregistrées de façon continue, en temps réel, et en tenant compte

des aspects quantitatifs et qualitatifs du comportement. La fidélité considérée sous ces aspects étant à peu près inexplorée, le deuxième objectif fut de faire l'inventaire des diverses approches de la fidélité[1] pour mieux éclairer sur les limites des méthodes appliquées et pour justifier la proposition d'une approche originale, surtout différente par le type de données traitées, et laquelle sera présentée en dernière partie.

Parmi les revues de littérature[2] ayant présenté un recensement des multiples études rapportant des méthodes d'évaluation de leurs données d'observation, on retrouve autant de façons d'aborder le sujet de la fidélité des données pour des raisons variées: soit à cause du domaine d'étude consulté (ex.: santé mentale, développement de l'enfant, adaptation scolaire, etc.), ou selon les méthodes d'enregistrement des données (enregistrement continu, par intervalles, etc.), ou enfin selon le milieu des recherches (études américaines, européennes, ou rattachées à des publications particulières). L'orientation choisie ici est arbitraire elle aussi, quoique nous ayons tenté d'être assez exhaustif quant aux ouvrages consultés.

Le processus d'analyse de la fidélité des données englobe de multiples considérations et leur spécification aidera à identifier le type de

---

[1] Il est à remarquer que les études de fidélité que l'on retrouve dans la littérature portent presque exclusivement sur l'observation appliquée; ceci permet de constater que les chercheurs sont davantage préoccupés à démontrer la validité externe de leurs découvertes d'observation qu'à rechercher des méthodes d'analyse statistique supportées empiriquement.

[2] Berk, 1979; Caro et al., 1979;Foster et Cone, 1980; Harris et Lahey, 1978, 1982; Hartmann, 1982; Hollenbeck, 1978; House, House et Campbell, 1981; Johnson et Bolstad, 1973; Kazdin, 1977;Keller;1980; Kelly, 1977; Kent et Foster, 1977; Kratochwill et Wetzel, 1977; Towstopiat, 1984; Yelton, Wildman et Erikson, 1977.

problématique commentée dans cette section. Ainsi, l'article publié par Hartmann (1982) fournit des explications détaillées sur les différents ordres de décisions impliqués dans ce processus: le chercheur doit d'abord déterminer les facettes (observateurs, système de codage, situations, milieux) du processus observationnel exigeant une évaluation formelle; il doit ensuite décider des conditions dans lesquelles les enregistrements de fidélité seront prélevés, choisir une unité d'analyse (grandeur et durée de la catégorie comportementale), sélectionner une statistique pour le sommaire de fidélité (procédure pour résumer la fidélité), interpréter les valeurs des statistiques de fidélité (juger de l'adéquation ou de la pertinence des données), modifier si nécessaire le plan de cueillette des données (moyens pour accroître la qualité des données), et enfin, rapporter les résultats de fidélité obtenus à partir des différents estimés dans des moments divers et sur chacune des variables étudiées. Dans cette longue énumération, la sélection d'une statistique destinée à résumer les aspects de fidélité représente le point d'intérêt de nos propos. Le choix d'une statistique de fidélité constitue, comme il a été révélé précédemment, une décision arbitraire et promise à la critique puisqu'elle ne peut reposer, à l'heure actuelle, sur des conceptions théoriques universelles. Les chercheurs disposent tout de même de certains critères pour soutenir leur choix et l'élucidation de ces assises décisionnelles a permis d'élaborer une nomenclature visant à distinguer les différentes approches d'analyse de la fidélité. Ainsi, cette section portera, en première partie, sur les critères de distinction des statistiques utilisées avec les données observationnelles; en deuxième partie, une classification des différentes approches d'analyse de

fidélité sera présentée; la troisième partie enchaînera avec la description des méthodes de calcul de la fidélité et de leurs variantes; finalement, la quatrième partie fournira quelques modèles d'interprétation des estimés de fidélité.

Critères de distinction

Les critères avec lesquels on peut établir des distinctions entre les diverses procédures d'analyse de la fidélité renferment des dimensions importantes que les investigateurs ont à considérer pour choisir la méthode statistique appropriée. Cependant, la référence à ces dimensions est rarement évidente dans les publications des travaux et ce n'est que dans les études de Hartmann (1977;1982), House, House et Campbell (1981) et Laurencelle (1981) qu'on les retrouve identifiées de façon explicite. Quatre principales considérations ressortent de ces études relativement au choix de la méthode d'analyse de la fidélité et elles seront résumées comme suit: (1) le type de données échantillonnées; (2) la contribution des jugements aléatoires au taux de concordance entre les observateurs; (3) les facteurs d'erreurs considérés; et (4) l'échelle de mesure de la statistique de fidélité visée.

A. Type de données échantillonnées

Principalement, deux types de données ont jusqu'à présent fait l'objet de la plupart des analyses de fidélité en recherche observationnelle sur le comportement humain. Ce sont les données numériques ou quantitatives, et les données catégorielles ou dichotomiques. Les données numériques proviennent

d'un échantillonnage continu d'événements ou d'intervalles temporels; elles fournissent respectivement des scores pour mesurer la fréquence et la durée d'apparition d'un comportement-cible. Hartmann (1982) explique que les scores quantitatifs représentent des réponses de fréquence, de taux, de temps de réaction ou de durée; ces mesures considèrent seulement le total des réponses enregistrées par les observateurs pour calculer leur concordance. Les statistiques corrélationnelles sont souvent utilisées pour résumer l'évaluation de fidélité sur ces mesures, dont les corrélations intraclasses qui définissent les rapports de variance entre les facettes étudiées. Dans les cas d'analyses corrélationnelles, les données sont organisées sur une échelle ordinale ou sur une échelle à intervalles. Toutefois, les mesures de pourcentages d'accords sont aussi calculées à partir des scores totaux enregistrés. Les données catégorielles[1] représentent des événements ponctuels codés de façon dichotomique (présence/absence, correct/incorrect, oui/non); elles font appel le plus souvent[2]à des procédures d'échantillonnage de temps. Cet échantillonnage comporte une segmentation préalable de la séance d'enregistrement en unités de temps égales; le profil des réponses enregistrées par deux observateurs sur la série d'intervalles ainsi constitués est présenté dans un tableau sommaire «2x2», lequel sera décrit plus loin. Puisque chaque

---

[1]Cette forme d'enregistrement des données fait l'objet de la majorité des travaux publiés relativement à l'observation directe sur le comportement humain.

[2]Le codage des catégories nominales par échantillonnage du temps d'observation est la forme d'encodage systématique la plus utilisée parce qu'elle simplifie les analyses des données. Toutefois, l'encodage systématique en temps réel est aussi possible lorsque les observateurs disposent de critères objectifs pour enregistrer les débuts et les fins des catégories nominales; le problème réside cependant dans la conception d'outils statistiques permettant l'analyse des données «flottantes», i.e. sans segmentation temporelle imposée.

type de données a ses méthodes d'analyses appropriées, les scores de catégories, ou scores nominaux, servent à évaluer la fidélité ou la précision inter-observateurs. Les mesures de pourcentage d'accords et leurs variantes constituent le choix le plus populaire pour les données catégorielles, même s'il est controversé.

D'autres aspects caractérisant la façon dont les données sont enregistrées (grandeur et étendue de l'unité de codage) sont aussi déterminants dans l'évaluation de la fidélité des données. Par exemple, un système d'enregistrement peut définir ses catégories de comportement de façon exclusive (une seule cote par objet d'observation: e.g.: «pleurer») ou de façon inclusive (des cotes multiples pour le même objet d'observation: e.g.: «comportement inadéquat» peut être codé par une ou plusieurs catégories complémentaires comme «frapper», «désobéir», «voler»); on dit alors que le niveau de mesure est à «cote simple» ou à «cotes multiples». Dans ce dernier cas, le calcul de la fidélité inter-observateurs est plus complexe et très peu d'études se sont penchées sur les problèmes d'analyses statistiques qu'elle comporte. Un autre aspect touchant à la grandeur de l'unité de score a trait à la multidimensionnalité du codage; ceci implique que l'objet d'observation soit enregistré dans plus d'une dimension, avec une liste de codes pour chacune. Le codage multidimensionnel complique aussi la méthode d'appréciation de la fidélité et peu d'analyses de la fidélité ont été publiées pour de tels systèmes

de codage[1]. Par rapport au dernier aspect, soit l'étendue du score, les mesures de fidélité porteront sur des <u>données par item</u> si le codage comporte une série d'enregistrements individuels comme dans le cas du codage par intervalles; si le codage est exécuté de façon globale sur une période plus longue, la mesure de fidélité est basée sur des <u>données de séance</u>. Cependant, ces deux types de mesures peuvent être rapportés pour une même étude. Passons maintenant à une deuxième considération permettant d'effectuer le choix d'une méthode statistique appropriée.

B. <u>Contribution du hasard à la concordance des observateurs</u>

Lorsque les chercheurs décrivent la fiabilité de leurs scores de catégorie par un relevé de la précision ou de la concordance des observateurs, ils négligent quelquefois de tenir compte des accords obtenus par le seul jeu du hasard. Toutefois, la nécessité d'apporter une correction pour la contribution du hasard au degré de concordance entre les jugements des observateurs sur ce type de données est grandement discutée (Fleiss, 1975; voir Hartmann, 1982). Il s'ensuit que les statistiques de fidélité des données catégorielles varient à la fois sur l'inclusion d'une correction, sur la façon d'inclure la correction et sur l'importance de la correction pour les accords imputables au hasard. Ainsi, le simple pourcentage d'accords (**(Accords/Accords + Désaccords) x 100**) n'en fournit pas; par contre le pourcentage d'accords des scores de catégorie calculé sur les seuls événements moins susceptibles d'être codés de façon

---

[1] Le lecteur pourra se référer à l'étude de Déziel (1985) pour un exemple d'analyses de fidélité inter et intra-observateurs à partir du codage multi-dimensionnel des mouvements du corps.

aléatoire (i.e. les événements moins nombreux: «**Accords sur événements présents**» ou «**Accords sur événements absents**») a un effet surcorrectif en omettant les autres accords. D'autres statistiques d'accords sur les scores de catégorie rejettent les accords anticipés par la chance seulement, et Hartmann (1982) cite le "kappa" de Cohen comme la plus populaire. Le kappa ainsi que d'autres méthodes de même type seront décrits plus loin. Parmi les nombreuses statistiques d'accords "2x2" utilisant des variables dichotomiques, Conger et Ward (1984) identifient seulement cinq indices différents relativement à l'inclusion d'une correction pour les accords imputables à la chance: modification du 'A', par Fleiss (1975); le coefficient phi (Ø); le kappa de Cohen; et deux coefficients de corrélation intra-classe. Ces auteurs concluent que ces indices sont fonctionnellement équivalents et que le chercheur peut baser son choix sur tout critère arbitraire. Ils fournissent des explications sur les possibilités et les limites de différents critères soustendant le choix d'un indice correctif pour mieux orienter les chercheurs dans cette décision. En résumé, les techniques d'estimation des probabilités d'attribution des cotes au hasard dont il a été fait mention s'adressent spécifiquement à des mesures de type présence/absence avec cotes simples; pour les mesures numériques, les statistiques corrélationnelles permettent d'évaluer l'importance des erreurs attribuées au hasard, soit les erreurs d'imprécision. Il sera donc expliqué plus loin comment une correction pour l'effet du hasard sur la concordance des résultats d'observation peut être inappropriée et comment certaines techniques sont inopérantes avec d'autres types de données. Nous verrons dans le prochain paragraphe que les différentes

façons d'interpréter les sources d'erreur représentent un autre critère distinguant les statistiques de fidélité.

C. Facteurs d'erreurs considérés

Un autre aspect à considérer lors du choix d'une statistique de fidélité réside, selon Hartmann (1982), dans le type d'erreurs qu'une statistique examine. Certaines statistiques traitent seulement les divergences dans la fréquence des réponses comme de l'erreur (e.g.: simple pourcentage d'accords basé sur les réponses totales), alors que d'autres considèrent en plus les divergences de localisation temporelle (e.g.: pourcentage d'accords sur les accords à chaque intervalle de codage). Hartmann spécifie que la plupart des statistiques corrélationnelles sont calculées sur les observations totales, alors que la plupart des mesures d'accord (incluant kappa, phi et les variantes du simple pourcentage d'accords) se basent sur des accords par item ("interval-by-interval agreements").

Hartmann poursuit sur une autre façon de traiter l'erreur dans les scores: les différences systématiques (biais) et les différences dues au hasard sont confondues ou, seulement les erreurs de hasard sont considérées comme une contribution à l'erreur. Pour cet auteur, les statistiques d'accords ("agreement") se rattachent à la première façon alors que les statistiques dites «de consistance» ("consistency") traitent uniquement les différences de hasard. Hartmann spécifie que les mesures de corrélations intraclasses utilisées dans l'évaluation de la fiabilité inter-observateurs des données quantitatives

peuvent être calculées soit avec les composantes de hasard seules, soit en ajoutant les différences systématiques entre observateurs comme contributions à l'erreur.

D. Echelle de mesure de la statistique de fidélité

Enfin, cette dernière considération mérite aussi attention lors de la sélection d'un indice approprié. La plupart des indices de fidélité prennent une valeur entre '0' et '1' mais la signification de cette valeur peut différer substantiellement à cause de l'échelle de mesure sous-tendant la statistique utilisée. Hartmann (1982) mentionne les trois échelles les plus utilisées: (1) les échelles de probabilité conditionnelle; (2) les échelles de proportion d'accords; et (3) les échelles de proportion de variances. Ces modèles d'interprétation des indices de fidélité seront explicités dans la quatrième partie de cette section,intutilée "Quelques modèles d'interprétation des indices de fidélité".

Hartmann (1982) ajoute d'autres critères de différenciation des statistiques sommaires de la fidélité, tels que: la capacité pour un seul indice à informer sur la valeur du système d'observation en entier (e.g.: accord brut et statistiques du type kappa) versus un calcul sur chaque facette d'observation procurant des scores (e.g.: corrélations intra-classes); leur aptitude à résumer plus de deux niveaux d'une facette (e.g.: plusieurs observateurs, essais ou sessions - par les corrélations intra-classes, et plusieurs statistiques d'accords, tel le kappa);leur relation formelle avec la théorie des tests ou avec

la théorie de la généralisabilité (e.g.: statistiques corrélationnelles incluant les corrélations intra-classes); et en dernier, leur degré de complexité et d'abstraction de calcul.

Avant de décrire les différentes méthodes d'évaluation de la fidélité déjà abordées dans cette section, une nouvelle nomenclature des approches de la fidélité est apportée dans la classification ci-après.

Classification des approches

Les différentes approches de la fidélisation des données d'observation directe puisent leur spécificité dans la nature de l'unité observable du système d'enregistrement ainsi que dans les objectifs particuliers d'une recherche. Ainsi, un chercheur peut désirer apprécier la fidélité de ses données pour de multiples raisons telles que: juger de la pertinence ou de l'ambiguïté des catégories comportementales qu'il a choisies; évaluer la performance de chacun des observateurs; connaître l'efficacité de l'entraînement des observateurs; comparer les sujets observés; obtenir une estimation globale de la valeur des observations enregistrées durant une ou plusieurs séances différentes. Idéalement, la fidélité des données devrait être établie sur l'unité de base faisant l'objet des observations avant d'être examinée pour une séance totale; par exemple, si le plan d'observation est conçu pour enregistrer cinq catégories comportementales dans une suite d'intervalles prédéfinis, on devrait en premier apprécier la fidélité pour les intervalles, ensuite pour chacune des catégories enregistrées et, après, pour l'ensemble de la séance. La procédure

logique dont il est ici question, soit l'évaluation de chaque unité et de chaque étape du processus observationnel avant l'estimation du processus entier, n'est pas toujours respectée. Ceci complique la comparaison des résultats de fidélité entre les études. Le répertoire des objets d'analyse de la fidélité, présenté ci-après, a été préparé dans le but de détailler et de distinguer les diverses étiquettes utilisées pour les mêmes approches. Cette classification sera suivie de trois tableaux résumant le matériel présenté dans les sections 'A' et 'B' et introduisant la section 'C'. Ces tableaux-résumés apparaissent paradoxalement comme une condensation des méthodes d'analyse de fidélité en observation directe, tout en montrant le foisonnement des méthodes.

A. Fidélité par événement séparé[1].

L'événement séparé se définit de deux façons:

1) Comme l'événement ponctuel[2] observable à chaque centration sur l'objet observé. Cette centration est organisée par un échantillonnage du temps d'observation en intervalles séparés de durées fixes; pour chaque intervalle, l'observateur enregistre par un code '1' (ou 'présent') si le phénomène-cible[3] a

---

[1]Le terme anglais "trial reliability" correspond à cette appellation; on peut aussi traduire par «fidélité d'items» ou «fidélité d'intervalles».

[2]Il est à noter que l'événement ponctuel est enregistré habituellement sous une seule cote; lorsque des définitions complémentaires s'appliquent pour un même événement, on aura des enregistrements variés (ou des cotes multiples) pour le même objet d'observation. Cette dernière procédure a peu été examinée sous l'angle de la fidélité (cf. Laurencelle, 1981).

[3]Lorsque, dans un même intervalle, plusieurs phénomènes-cibles doivent être repérés, l'observateur enregistrera plusieurs codes en fonction de toutes les catégories de comportement observées. C'est le cas pour l'observation multidimensionnelle, ou pour l'enregistrement avec cotes multiples. Cependant, les méthodes d'analyse de la fidélité qui ont été transposées de la théorie classique des tests ne sont guère applicables dans ces cas.

pu être observé et par un code '0' (ou 'absent') s'il n'est pas apparu. On obtient ainsi des données dichotomiques de catégorie.

2) Comme chaque centration successive, sans échantillonnage du temps, où on enregistre alors des états par leur durée ou leur fréquence; ce sont des données numériques.

La fidélité par événement séparé a pour but d'indiquer l'adéquation de la définition du comportement observable ainsi que l'aptitude de l'observateur à utiliser cette définition et tout le matériel d'observation. Une telle évaluation peut donc aboutir à une redéfinition de l'objet d'observation pour le rendre mieux observable, ou à un réentraînement de l'observateur en vue d'améliorer sa performance.

Kent et Foster (1977) identifient quatre types de mesures de fidélité par événement séparé: (1) fidélité d'accords; (2) fidélité d'événements présents ou d'événements absents (items codés ou non-codés); (3) fidélité corrigée pour la chance sur les événements présents ou sur les événements absents (un seul type d'événement est retenu, soit le moins fréquent); (4) fidélité d'accords corrigée pour les accords imputables au hasard (ex.: mesures du type Kappa de Cohen). Hartmann (1977, 1982) fait simplement la distinction entre les statistiques d'accords et les statistiques de type corrélationnel. Toutefois, les statistiques pour la fidélité des données par événement séparé ("trial reliability") diffèrent selon que les données sont «catégorielles» ou «numériques». Hartmann (1977) explique que les données numériques ou

quantitatives sont analysées avec les mêmes statistiques que celles employées dans la fidélité par séance ("session reliability"). Ces méthodes différentes seront explicitées plus loin.

B. Fidélité par séance[1] (ou par situation)

La séance est la période d'observation spécifique à un moment ou à un lieu, et choisie par le chercheur (e.g.: repas du soir, période de jeu libre, dix premières minutes d'une entrevue de thérapie, etc.). Elle peut être constituée de la somme des événements séparés échantillonnés systématiquement (données de présence et d'absence: e.g.: l'addition des scores de 20 périodes d'enregistrement de 15 secondes quotidiennes pour le comportement d'«attention» d'un étudiant ou l'addition des temps de réaction pour chacune des réponses aux 30 requêtes quotidiennes faites par un enfant), ou du total des incidents observés de façon continue (données de fréquences et de durées: e.g.: compter le nombre d'incidents d'aide durant une période de 20 minutes de jeu libre). Les évaluations de la fidélité par séance servent à comparer les scores totaux entre les observateurs pour être en mesure ultérieurement de porter un jugement sur la valeur des résultats d'observation. Hartmann (1977) spécifie que la fidélité par séance indique le degré de généralisation des observations (ou scores de séance) d'un juge aux observations qu'un autre juge pourrait obtenir. Puisque les scores de séance sont des scores composés de multiples événements réels ou artificiellement séparés, ils seront habituellement plus

---

[1] La documentation américaine y réfère par "session reliability" ou "total reliability", ou encore "frequency agreement".

fidèles que les scores des composantes. Les mesures de fidélité par séance peuvent représenter des moyennes, des totaux ou des échelles de temps; elles varient de zéro à une valeur positive quelconque et peuvent être considérées comme ayant les propriétés d'une échelle de proportion. Les statistiques utilisées sont le pourcentage d'accords global et, plus souvent, les méthodes corrélationnelles. Toutefois, pour une séance comportant des cotes multiples pour plusieurs catégories, ou pour plusieurs dimensions, les méthodes corrélationnelles ne conviennent pas et les méthodes de pourcentages nécessitent plus de raffinement.

C. Fidélité par événements continus multiples

Les événements continus sont une suite d'intervalles inégaux, ou états, notés par leur début et par leur fin durant une période d'observation donnée. Il y a deux niveaux d'enregistrement des données: pêle-mêle ou systématique. Dans la majorité des études observationnelles, cette approche de fidélité est considérée comme une fidélité par séance parce que les états sont analysés de façon unidimensionnelle, i.e. comme si une seule dimension avait été observée. On utilise alors les mêmes statistiques que pour la fidélité par séance. Cependant, quelques études se sont penchées sur le problème de calculer la fidélité des états qui se chevauchent à cause de la multidimensionnalité du codage. Les méthodes d'analyse de la fidélité dans cette optique sont devenues complexes à cause des différences de segmentation temporelle des données d'un observateur à l'autre.

D. Fidélité globale

Différentes utilisations des scores mènent à une mesure de fidélité globale. Quelquefois les scores de séance sont utilisés pour obtenir un indice de fidélité globale. Il est cependant nécessaire de distinguer «mesure de séance» et «mesure globale» puisque la dernière s'applique dans le sens de valeur globale pour tout un processus. Un processus général contient plusieurs mesures de plusieurs séances. Les méthodes d'analyse de la fidélité globale sont souvent les mêmes que pour la fidélité par séance. Une technique plus appropriée serait d'apprécier la fidélité pour chaque événement séparé ou chaque catégorie et de calculer la moyenne des différents indices obtenus (voir Laurencelle, 1986).

E. Fidélité par catégorie

Une catégorie a trait à chaque définition comportementale qui se résume par une cote nominale ou par un code. On peut aussi avoir des valeurs numériques pour une catégorie lorsque l'on s'intéresse à la durée ou à la fréquence d'une catégorie nominale. En fait, on peut faire une analyse de la fidélité par catégorie à partir des trois formes d'encodage des données (par intervalle, par séance et de façon continue). L'encodage de façon continue, en temps réel, donnera cependant une meilleure appréciation de la précision des catégories puisque, contrairement aux autres formes d'encodage, il ne résulte pas d'une approximation des catégories apparues. Les indices statistiques utilisés offriront des mesures de concordance entre les paires de codeurs. En

interprétant la valeur de la concordance inter-codeurs ou intra-codeur pour une catégorie, on veut donc se prononcer sur le degré de clarté d'une description comportementale, i.e. sa facilité à être reconnue sans ambiguïté. Par exemple, après une analyse de fidélité par catégorie, on peut découvrir qu'une catégorie est peu utilisée, ou qu'une autre est majoritairement choisie, qu'une catégorie obtient une très faible proportion d'accords, que certaines paires de codeurs s'entendent mieux sur telles catégories en particulier, ou encore qu'un des codeurs dévie constamment dans ses choix de certaines catégories; la liste pourrait s'allonger d'exemples propres à chaque étude. De plus, comme il a été mentionné auparavant, une catégorie peut être enregistrée selon deux principes différents: exclusif ou complémentaire. Ce dernier aspect implique des difficultés au niveau de la conception des outils statistiques, de même qu'il oblige le chercheur à raffiner ses définitions comportementales.

F. Fidélité par observateur

Cette fidélité implique l'analyse spécifique de la performance des observateurs. Dans ce type d'étude, le chercheur se questionne sur la fiabilité des observateurs. Elle peut être abordée sous divers aspects: apprécier la cohérence d'un codeur avec lui-même, déterminer la variation de l'accord parmi les paires de codeurs, ou comparer les cotes des observateurs avec des cotes standard. Le point d'intérêt est donc uniquement d'obtenir des estimés du rendement des observateurs.

G. Fidélité par sujet observé

Dans la recherche d'observation appliquée, on veut interpréter à quel point les données reflètent les vrais comportements des sujets. Cette question requiert que le même sujet soit observé par différents observateurs et à différents moments. Ainsi, on suppose que plus les données concordent entre les observateurs, plus il y a de chance que les variations observées chez les sujets soient réelles. Les méthodes corrélationnelles intraclasses sont souvent recommandées pour ce type d'analyse.

Tableau 2

Forme des analyses déterminée par le choix du **principe de codage**, des **critères de cotation**, et du **type d'observations**

Type d'encodage & données	Critères de cotation	Type d'observations
Par <u>blocs de temps</u> Données nominales (présence/absence) <u>Centrations</u> selon occurrences des comportements Données nominales <u>Continu et complet</u> Données nominales et données continues (à intervalles, ordinales,etc.)	<u>Cotes</u>: simples/ multiples <u>Durée</u>: fixe/ variable <u>Catégories</u>: unidimensionnelles/ multidimensionnelles	<u>Fréquence</u>: nombre de fois qu'une catégorie a été choisie (ou qu'un comportement a été repéré) <u>Durée</u>: longueur totale ou moyenne de l'occurrence temporelle d'une catégorie ou d'un ensemble de catégories <u>Séquence</u>: ordre sériel ou chronologique d'un ensemble de catégories selon certains caractères déterminants <u>Intensité</u>: degré d'amplitude d'un phénomène

Tableau 3

Approches de fidélité en fonction des objectifs d'évaluation des données

Approches de fidélité	Appréciation de la valeur des résultats	Objectifs d'évaluation
Par événement séparé	Pour chaque observation	Ambiguïté de chaque observation
Par séance	Pour chaque période (ou chaque séance)	Valeur globale d'une séance choisie
Par événements continus multiples	Pour chaque ensemble d'observations multidimensionnelles ou chaque segmentation temporelle	Correspondance des états qui se chevauchent ou qui s'enchaînent
Globale	Pour le processus observationnel en entier (observateurs, données, méthode)	Valeur globale du processus observationnel
Par catégorie	Pour chaque description comportementale	Clarté des définitions comportementales
Par codeur	Pour chaque codeur avec lui-même ou avec un critère	Performance moyenne d'un codeur, qualité de l'entraînement
Par sujet observé	Pour chaque sujet de l'expérience par rapport à une catégorie spécifique ou à un ensemble de catégories	Applicabilité des catégories du répertoire aux sujets choisis

Tableau 4

Principaux indices statistiques de fidélité

Données	Indices
<u>Nominales</u>:	Indices se basant sur le tableau «2 x 2»: - Pourcentage d'accords et variantes - Kappa* - Phi** (indice corrélationnel pour données dichotomiques avec une seule catégorie) - Pi* de Scott, Lambda*, Chi-carré Pourcentage d'accords global (pour chaque item ou pour tous les items)
<u>Numériques</u>:	Pourcentage d'accords global (sur les durées) Indices corrélationnels**:
ordinales	- Corrélation de rang de Spearman - W de Kendall
intervalles	- Corrélation de Pearson - A de Robinson - Corrélation intraclasse (coefficients de généralisabilité)

* Ces indices apportent une correction pour la concordance imputable au hasard.

** Ces indices comportent un contrôle pour la concordance imputable au hasard, selon House, House et Campbell (1981).

Description des méthodes d'évaluation de la fidélité

En définitive, un grand nombre de chercheurs en observation appliquée rapportent des indices de fidélité avec le seul souci de montrer que leurs études ont été exécutées selon les conventions métriques générales. Les motifs de ces recherches étant avant tout de faire des conclusions à partir des observations recueillies, on n'accorde que peu d'intérêt à questionner la pertinence des analyses de fiabilité des données. Donc, les approches choisies pour rendre compte de la valeur des données, soit la «fidélité par événement séparé», la «fidélité par séance» ou quelque autre, sont davantage des décisions accidentelles plutôt que planifiées; c'est-à-dire que la structure des enregistrements recueillis conditionne le choix de l'approche. Idéalement, une étude de fidélité bien conduite devrait apprécier le processus observationnel en entier (unités d'enregistrement, catégories comportementales, séances, observateurs, sujets); par conséquent, elle comporterait différentes étapes où des approches particulières seraient indiquées. Cependant, la procédure actuelle est généralement de développer des méthodes de cueillette des données simplifiées pour ne pas affronter les complexités statistiques. Ceci explique que la majorité des recherches choisissent l'enregistrement par échantillonnage de temps menant à des données dichotomiques regroupées dans un tableau «2x2» pour les fins d'analyse de la concordance des observateurs. Ainsi, dans la section qui suit, les méthodes d'évaluation de fidélité qui sont relatées concernent en grande partie des approches de fidélité par événement

séparé et par séance puisqu'elles s'appliquent à des données catégorielles et dichotomiques.

1. Méthodes de pourcentages. Puisque les méthodes de pourcentages concernent plus souvent des données nominales, elles seront énumérées séparément des méthodes corrélationnelles, ces dernières s'appliquant aux données quantitatives. Parmi les méthodes de pourcentages, il est indispensable de distinguer les mesures d'accords se basant sur les tableaux «2x2» des autres mesures donnant plus de nuances aux incidences d'accords ou de désaccords. Le regroupement des méthodes sous cinq étiquettes différentes est justifié par le seul objectif de rendre ces distinctions plus évidentes.

a. Le pourcentage d'accords simple. Le pourcentage d'accords global est l'indice le plus simple servant à déterminer l'accord inter-juges sur les scores totaux de séance. Hartmann (1977) identifie deux variantes[1] du pourcentage d'accords global: (1) Le premier indice est calculé sur les fréquences d'occurrences seulement; on divise le plus petit score d'occurrences de deux observateurs, pour une séance, par le plus grand score, et on multiplie par 100. (2) Le deuxième indice tient compte, en plus des fréquences, de l'intersection des occurrences codées; on divise le nombre d'occurrences où il y a croisement entre les deux observateurs par le nombre d'occurrences différentes réunies pour cette même séance, et on multiple par 100. Par exemple, pour les deux ensembles de scores suivants, "0 1 0 0 1" et "0 0 1 0 1",

[1] Ces calculs du pourcentage d'accord global s'appliquent aux données d'une seule catégorie de comportement.

le premier indice présente un accord parfait (100%) et le deuxième indice donne un résultat d'accord de 33%.

(F5)

1) $\%A = \frac{\text{Min. (\# occurrences)}}{\text{Max. (\# occurrences)}} \times 100$

2) $\%A = \frac{\text{\# occurrences en intersection}}{\text{\# occurrences en union}} \times 100$

Hartmann (1977) explique que la première variante du pourcentage d'accords global est attirante par son utilité à évaluer si la différence entre les scores de séance représente un vrai changement, ou simplement l'erreur d'observation. Cependant, il précise qu'elle comporte de nombreuses limites dont l'absence d'une borne inférieure d'acceptabilité et d'une valeur indiquant l'absence d'accord. Sa valeur statistique est dépendante du taux de production des comportements, ou du taux d'apparition des catégories pour une séance donnée. Ainsi, une fréquence relativement plus élevée d'un ou de quelques comportements engendrera un nombre plus élevé d'accords.

Quant à la deuxième variante du pourcentage d'accords global, Hartmann (1977) prétend qu'elle présente une appréciation très sévère, utilisant peu l'information présente dans les données. Pour leur part, Hawkins et Dotson (1975; voir Kratochwill et Wetzel, 1977) critiquent la formule générale «**(Accords/#items)x100**» (soit la deuxième variante du % d'accords global) en mentionnant qu'un simple jugement sur la compétence des observateurs à

partir de leurs observations concordantes n'est pas suffisant pour tirer des conclusions à propos de l'objet d'étude. Le seul résultat de la proportion des concordances sur le total d'observations ne peut donc pas mener à un questionnement sur l'adéquation des définitions comportementales, le profil des désaccords n'étant pas examiné.

b. Les mesures d'accords basées sur le tableau «2x2». La méthode la plus populaire pour calculer la fidélité des données d'intervalles traitant une seule catégorie à la fois, s'énonce comme la formule du pourcentage d'accords simple :

(F6)

$$\%A = [Accords/(Accords + Désaccords)] \times 100$$

Cependant, cette formule réfère au tableau sommaire regroupant les incidences d'accords (cases A et D) et de désaccords (cases B et C) entre deux observateurs pour une catégorie donnée de comportement. House, House et Campbell (1981) illustrent ce tableau ainsi:

Tableau 5

Illustration du TABLEAU «2x2»

		Observateur 2		
		***Présence*(+)**	***Absence*(-)**	
	Présence (+)	**A** (oui/oui)	**B** (oui/non)	A + B*
Observateur1	***Absence*** (-)	**C** (non/oui)	**D** (non/non)	C + D*
		A + C*	B + D*	

*= totaux marginaux

Le nombre d'intervalles codés équivaut à la somme des cases **A+B+C+D**. La fréquence d'occurrences codées est obtenue en additionnant les cases A et B pour l'observateur 1 et les cases A et C pour l'observateur 2. Le pourcentage d'accords des données échantillonnées par intervalles oblige à ce qu'un tableau «2x2» soit produit pour chacune des catégories comportementales du répertoire.

House, House et Campbell (1981) ont dénombré six formules de pourcentages d'accords qui sont des variations de la formule générale se basant sur les cases du tableau «2x2». Elles diffèrent selon l'importance qu'elles

assignent à certaines cases[1]. Le cas extrême est d'assigner '0' à une case pour éliminer son influence sur la mesure d'association (e.g.: **[A/(A+B+C)] x 100** ou **[D/(B+C+D)] x 100**). Cette méthode de calcul exclut les contributions de la case avec un nombre élevé d'accords, lesquels peuvent être attribuables en grande partie à la chance. La part de la chance consiste dans le nombre anticipé ou la proportion d'accords qu'on obtiendrait lorsque deux observateurs ont une grande probabilité d'arriver à des jugements communs sur un nombre de cotes ou d'items donnés; il s'agit en fait des évaluations d'un observateur (ou des deux observateurs) qui sont sans relation avec la connaissance du fait observé. Pour les tableaux 2 x 2, les accords anticipés en vertu du hasard sont calculés à partir des valeurs marginales comme suit: **[(A+B)(A+C)/N] + [(B+D)(C+D)/N]**, '**N**' étant le nombre total d'intervalles observés. Toutefois, le calcul du nombre d'accords anticipés ne détermine pas quelle est la proportion des accords réels, ou quel est le taux d'accords vraiment attribuable au hasard; il indique seulement la probabilité maximale d'accords aléatoires avec un profil spécifique des accords et des désaccords entre deux observateurs. Hartmann (1977) explique que d'autres évaluations de la fidélité sont requises pour déterminer le taux d'accord réel.

Le pourcentage d'accords calculé à partir du tableau «2x2», est grandement influencé par le choix du type d'incidences d'accords (événements

---

[1]Certaines études appliquent la formule du pourcentage d'accords en se basant seulement sur les accords d'événements présents (case A) ou d'événements absents (case D); c'est la méthode identifiée «accords pour événements spécifiés» ou «pourcentage d'accords effectifs». La méthode donnant un pourcentage d'accords total utilise la somme des accords sur les deux types d'événements, i.e. **[(A + D)/(A+B+C+D)] x 100**.

présents ou absents, ou les deux). Par exemple, pour une séquence de 100 intervalles observés, avec 15 incidences d'accords sur la présence de l'événement, 80 accords sur l'absence et 5 désaccords, on obtient respectivement 75%, 94% et 95% selon que l'on a opté pour la formule basée sur les événements codés, non-codés ou les deux. Aussi, la fréquence d'apparition des événements-cibles influence la probabilité d'accords sur les présences ou sur les absences d'événements en produisant un taux de concordance gonflé, lequel ne montre pas l'accord réel entre les observateurs. Cette situation se retrouverait, par exemple, dans le cas où deux observateurs enregistrent un comportement présent dans 10% des intervalles en ne s'entendant pas sur les moments où il apparaît (case A = 0, case B = 10, case C = 10 et case D = 80); le 80% d'accords sur les non-apparitions du comportement ne fournira pas d'information quant à la difficulté de reconnaître le phénomène étudié. Pour susciter une investigation plus poussée, il serait plus utile de montrer la non-concordance des juges quant à l'apparition du phénomène. Cet exemple révèle l'importance de choisir la formule de calcul donnant un indice d'accords inter-observateurs qui montre le mieux l'état réel de la concordance par rapport au phénomène étudié. Dans le cas inverse, où deux observateurs enregistreraient chacun la présence d'un comportement dans 90% des intervalles (case A = 80, case B = 10, case C = 10 et case D = 0), en ne s'entendant pas sur 10% de ces présences, le 80% d'accords apparaît trop facilement obtenu; le chercheur aurait avantage à s'interroger sur la pertinence de ce résultat. Comment le hasard a-t-il influencé les choix de réponses des observateurs? La définition du comportement à repérer est-elle trop générale? Si la probabilité que le

comportement apparaisse était très élevée, alors les juges n'ont pas vraiment eu à faire un choix à chacun des 80 intervalles codés. Le désavantage des données enregistrées par intervalles réside dans la difficulté de déterminer le nombre de comportements différents; ainsi, un pourcentage d'accords élevé comme celui ci-dessus est relatif puisqu'il peut être à la fois basé sur un comportement bref s'étant répété sur plusieurs intervalles et sur un même comportement se poursuivant durant plusieurs intervalles. Le tableau «2x2» n'est donc pas conçu pour nuancer les résultats d'accords puisqu'il se base sur des approximations de la fréquence des comportements.

House, House et Campbell (1981) font mention du manque de rigueur de la formule du pourcentage d'accords total, calculé à partir du tableau «2x2» **[(A+D/A+B+C+D)x100],** lorsqu'elle est appliquée aux comportements avec une fréquence d'apparition en dehors de la marge modérée (40-60%). L'approche conservatrice de restreindre l'attention à un type d'événement (présences ou absences) a aussi un inconvénient: celui de sous-estimer l'accord dans le cas d'événements très peu fréquents. Ainsi, les mesures de pourcentages tirées de ce tableau sont dépendantes de la grandeur du nombre d'intervalles pertinents, ou codés. Harris et Lahey (1978) admettent que la méthode restreignant le calcul de l'accord inter-observateurs à un type d'événements (choix des intervalles codés (case A) si le taux des comportements observés est faible, ou choix des intervalles non-codés (case D) si le taux des comportements observés est élevé) réduit le biais introduit pour de tels accords imputables au hasard. Ils ajoutent cependant qu'elle comporte plusieurs lacunes comme la

surcompensation des accords d'un type d'événements au détriment de l'autre, l'inadéquation du pourcentage d'accords lorsque le taux des comportements codés varie, ainsi que l'absence de règles objectives déterminant le choix particulier d'une fréquence de scores.

Dans leur revue de 1981, House, House et Campbell mentionnent que les variations entre les formules de pourcentages d'accords «2x2» proviennent des interprétations différentes sur ce qui devrait être compté comme «accord» et ce qui devrait être compté comme «désaccord». Ils commentent les avantages et désavantages ainsi: la facilité de calcul et d'interprétation constitue le seul avantage majeur; le seuil d'accord convenable est indéfinissable; un consensus entre chercheurs laisse croire que le 70% est nécessaire, le 80% est adéquat et le 90% est bon; l'absence d'un facteur de correction pour les différences de cotes attribuables à l'effet du hasard représente la limite la plus pénalisante des formules d'accords «2x2»; l'effet de hasard peut se répercuter sur l'accord entre observateurs à différents niveaux de la fréquence des réponses comportementales et la probabilité d'accords dus à la chance diffère pour chacun; un biais s'introduit dans la mesure si cette correction n'est pas apportée.

c. Méthodes «2x2» améliorées. Plusieurs arrangements ont été proposés relativement aux méthodes précédentes basées sur les données placées en tableau «2x2» et ils sont passé en revue par House, House et Campbell (1981). La formule de Hawkins et Dotson (1975) offre un compromis à la méthode retenant l'accord des événements codés seulement:

(F7)

$$\text{\% moyen d'accords sur événements présents et absents} = \frac{[A/(A+B+C)] + [D/(B+C+D)]}{2} \times 100$$

Farkas (1978) tente de remédier au biais possible apporté par l'inégalité du nombre d'intervalles impliqués dans les deux composantes de la formule précédente en proposant un pourcentage d'accords total pondéré, lequel s'obtient par: **[(A+D)/(A+D+2(B+C))] x 100**. Sloat (1978) critique cette correction en démontrant qu'elle ne fait qu'ajouter un facteur additionnel de pénalité pour les erreurs. Une autre mesure pondérée du pourcentage d'accords «2x2», cette fois-ci basée sur les accords d'événements présents, utilise un des deux observateurs au statut «critère» et traite les erreurs d'omissions du second observateur par rapport aux événements codés par l'observateur «critère»: **(A/A+B) x 100** ou **(A/A+C) x100**. Le choix de statut des observateurs y est déterminé par le chercheur et non par les données obtenues. La rapidité et la facilité de calcul semblent être les seules justifications pour ignorer la source d'un type de désaccords.

Dans une autre approche[1] du problème des différences dans l'accord dû au hasard, différences causées par les nombres inégaux des intervalles codés et non-codés, on assume que l'accord dû au hasard est moins probable lorsque l'on prend en considération l'événement le moins fréquent (présence ou absence).

---

[1] Par Clement (1976; voir House, House et Campbell, 1981; Harris et Lahey, 1978).

L'auteur prend un observateur comme «critère», choix qui peut être fait au hasard, et obtient la formule qui suit:

(F8)

$$[(A \times B)+(C \times D)]$$

$$A = \frac{\text{Accords sur événements présents}}{\text{\#événements codés par observateur standard}} \qquad B = 1.00 - \frac{\text{\#événements présents de l'observateur standard}}{\text{\# total d'échantillons de temps}}$$

$$C = \frac{\text{Accords sur événements absents}}{\text{\#événements non-codés par observateur standard}} \qquad D = 1.00 - \frac{\text{\#événements absents de l'observateur standard}}{\text{\#total d'échantillons de temps}}$$

Ce calcul procure une moyenne équilibrée d'indices d'accords sur événements présents et absents avec une pondération assignée à ces deux indices selon la fréquence à laquelle le comportement est enregistré. Son avantage est de compenser pour les distorsions dues aux accords de chance des fréquences élevées ou faibles sans avoir à éliminer de données. Harris et Lahey (1978) ont apporté deux modifications à la formule de Clement: (1) les diviseurs des termes A et C prennent les valeurs d'accords ou de désaccords sur les événements présents ou les événements absents; (2) les termes B et D deviennent des proportions moyennes d'intervalles codés ou non-codés, divisées par deux. La formule se conceptualise comme de l'accord d'événements présents équilibré par le taux moyen d'événements absents, plus de l'accord d'événements absents équilibré par le taux moyen d'événements présents. Le

facteur d'équilibration est la moyenne enregistrée par les deux observateurs, plutôt que d'en désigner un arbitrairement comme le proposait Clement.

Yelton et ses collaborateurs (1977) ont conçu une statistique traitant la probabilité que le nombre d'accords observés, ou plus, soit obtenu uniquement par la chance. Leur formule utilise des expressions factorielles et, selon House, House et Campbell (1981), sa complexité de calcul n'en fait pas une approche réaliste et appréciable. Harris et Lahey (1978) considèrent que cette méthode ne décrit pas le degré d'accord mais qu'elle procure seulement une comparaison entre le score d'accords obtenus et le score d'accords prévus. Birkimer et Brown (1979) ont utilisé la formule de Yelton et ses collègues (1977) en y ajoutant une analyse graphique des pourcentages de désaccords obtenus dans chaque condition des relevés d'accords inter-observateurs. Le but visé par la présentation d'un support graphique est de juger à quel point les données obtenues avec un accord réel entre observateurs sont pertinentes avec les variables étudiées. Leurs tableaux d'évaluation des niveaux de désaccords servent à juxtaposer les listes de taux de désaccords obtenus dans différentes occasions d'observation; si des taux et des types de désaccords se répètent dans différentes occasions d'observation, les conclusions sur les variables comportementales étudiées devront alors être formulées avec prudence. Ils suggèrent d'utiliser au moins 50 occasions d'observation et de calculer les pourcentages de désaccords des comportements rapportés à des taux de production moyens. Birkimer et Brown obtiennent une règle leur permettant d'évaluer si le taux de désaccords, ou d'accords, est simplement attribuable à la chance. Yelton (1979) propose l'utilisation de la règle de Birkimer et Brown

comme critère à rencontrer pour démontrer que le taux de désaccords entre observateurs est plus bas que ce qui peut être prévu par le hasard. Yelton effectue cette évaluation de la non-concordance entre observateurs pour nuancer ses résultats provenant d'une étude de la variabilité des données observées. Hopkins (1979) recommande la prudence sur l'adoption des conventions proposées par Birkimer et Brown. Il base sa critique sur le fait que l'augmentation des intervalles de vérification n'apporte pas plus de précision à la technologie d'observation même si cela porte le pourcentage de désaccords à un niveau plus acceptable. C'est pour lui une façon de rendre acceptable une faible fidélité observationnelle sans apporter d'amélioration à la technologie d'observation.

Dans leur étude de 1977, Kratochwill et Wetzel étiquettent les indices statistiques et graphiques représentant l'accord entre observateurs d'«aides» ou de «supports à juger» ("judgmental aids") de la crédibilité des données d'observation. Pour eux, le relevé ou le sommaire d'étalonnage de l'accord comporte des avantages et des limites: dans certaines conditions, les supports graphiques fournissent une source supplémentaire d'information (e.g.: détection des comportements critiques, maintien de contact avec les données, évaluation des différences absolues entre observateurs, détection de certaines menaces à la validité interne); quant à certains supports statistiques, comme ceux provenant des tableaux «2x2», ils peuvent augmenter artificiellement la crédibilité des données d'observation. Ces auteurs sont de ceux qui suggèrent

les statistiques du type Kappa et du type Phi comme des alternatives aux statistiques conventionnelles du pourcentage d'accords brut.

d. Mesures d'accords avec correction spécifique pour l'accord dû au hasard. Beaugrand (1982) fournit un tableau des principaux indices permettant d'estimer la fidélité appariés aux différentes formes de données. Pour les données de type nominal, il commence par exclure le pourcentage d'accords inter-juges dans sa forme la plus simple **(Acc./Acc.+Désacc.)** à cause qu'il ne tient pas compte des cas attribuables au hasard. Il souligne que l'indice Kappa de Cohen est le plus recommandé pour les systèmes à cotes nominales. La formule du kappa s'articule comme suit:

(F9)

$$\text{kappa} = \frac{(P_o - P_c)}{(1 - P_c)}$$

[P_o = la proportion des accords observés
P_c = celle des accords dus au hasard.]

Les accords pour chacun des comportements observés par deux observateurs sont placés sur une matrice et se situent sur la diagonale principale, alors que les désaccords se retrouvent de chaque côté de cette dernière. Le pourcentage d'accords observé (P_o) s'obtient en divisant le total des accords obtenu sur la diagonale par la sommation des événements observés, soit les totaux des accords et des désaccords sur les lignes et sur les colonnes. La somme des probabilités d'accords attribuables au hasard (P_c) s'obtient par l'addition des produits croisés des proportions relatives calculées aux lignes et aux colonnes. Le kappa est interprété comme la proportion de jugements

communs dans laquelle il y a accord après que l'accord imputable au hasard ait été exclu. La valeur de 'k' ne reposant que sur la comparaison entre l'ampleur de l'accord observé et l'ampleur de l'accord attendu par la chance seulement, nous obtenons évidemment un calcul sévère de la proportion d'accords lorsqu'il y a des désaccords. Par exemple, prenons les cinq profils de résultats suivants, basés sur des tableaux «2x2»:

(1)		(2)		(3)		(4)		(5)	/100 unités
A=70	B=10	A=70	B=10	A=70	B=5	A=70	B=5	A=70	B=0
C=10	D=10	C=5	D=15	C=5	D=20	C=0	D=25	C=0	D=30
kappa=0.38		k=0.57		k=0.73		k=0.88		k=1.0	

Ces profils montrent non seulement l'accentuation de la correction avec l'accroissement des désaccords entre observateurs, mais aussi une absence de correction lorsque l'accord est parfait (k=1.0). Une fois admise la possibilité que les observateurs s'accordent sur une base de hasard, pourquoi n'en pas tenir compte dans le cas d'un pourcentage d'accord parfait?

Divers autres profils de résultats placés en tableau «2x2» (voir schémas sur la page suivante) font apparaître certaines autres limites du kappa: par exemple, le coefficient kappa prend des valeurs négatives lorsque "BC > AD"; cette situation se produit lorsqu'un '0' est enregistré à la case 'A' ou 'D'. De plus, on obtient un coefficient '0' lorsqu'un accord parfait est obtenu sur les occurrences uniquement (case 'A'). Enfin, la valeur du kappa s'infléchit rapidement vers le bas lorsque le nombre d'intervalles codés diminue même légèrement. Ce comportement fait du kappa un indice descriptif peu

intéressant vis-à-vis du simple pourcentage d'accords, puisqu'il s'en écarte exagérément et ne semble plus refléter les données.

/90 unités /85 /80 /75 /70 Ex. (1)

A=70	B=10	A=70	B=10	A=70	B=5	A=70	B=5	A=70	B=0
C=10	D=**0**	C=5	D=**0**	C=5	D=**0**	C=0	D=**0**	C=0	D=**0**
k=-0.11		k=-0.08		k=-0.07		k=0		k=0 %A=1.0	

/100 unités /98 /95 /93 /90 Ex. (2)

A=70	B=10	A=70	B=10	A=70	A=10	A=70	B=10	A=70	B=10
C=10	D=10	C=10	D=8	C=10	D=5	C=10	D=3	C=10	D=0
k=0.38		k=0.33		k=0.22		k=0.08		k=-0.11	

Une extension du kappa a été formulée pour s'adapter aux accords simultanés entre plus de deux observateurs, et une pondération de l'indice permet d'identifier à l'avance que certains types d'erreurs sont plus importants que d'autres. Fleiss (1971) stipule que le kappa de Cohen (1960) est utilisé si tous les désaccords sont également sérieux; l'extension du Kappa (Cohen, 1968) sert lorsque l'importance relative des différents types de désaccords peut être spécifiée. Kent et Foster (1977) ont donné des formules dérivées, soit le kappa pour événements présents (K_o - "occurrence kappa") et le kappa pour événements absents (K_{no} -"nonoccurrence kappa").

House, House et Campbell (1981) mentionnent que les indices comme le **pi** (π) de Scott, ou le $r_{\pm}$ de Fleiss (dans les cas où les distributions marginales sont bien balancées), et le **lambda** sont aussi dérivés de la même formule théorique. Hollenbeck (1978) cite le **Pi** de Scott comme le premier coefficient d'accord présenté pour les échelles nominales et faisant une correction pour les

accords attribuables à la chance. Le **lambda** est semblable au 'K_0' de Kent et Foster. Toutes ces mesures prennent des valeurs entre +1 (accord complet) et 0 (désaccord complet). L'association complète est atteinte seulement lorsque les deux cases de fréquences d'erreurs du tableau sommaire (B et C) égalent '0'. Hollenbeck (1978) mentionne en plus la mesure du **Chi-carré** (χ^2) pour les échelles nominales; cette mesure fournit un index global d'association pour la distribution totale des codes observationnels; selon lui, elle est moins utile que le kappa parce qu'elle mesure l'association et non l'accord des codes. Par contre, Hollenbeck (1978) recommande le kappa non seulement à cause de sa facilité de calcul, mais à cause de ses propriétés métriques et de son lien mathématique avec les statistiques d'association.

Towstopiat (1984) critique les limites du kappa et de ses dérivés: l'indice kappa n'a pas une signification directement interprétable; les mesures kappa comparent le niveau d'accord global observé avec le niveau global attendu; elles n'évaluent pas les différences entre les patrons d'accords observés et attendus; le désaccord consistant entre observateurs ne peut être détecté et les catégories d'accords statistiquement différentes ne peuvent être déterminées car le kappa n'offre pas la possibilité de partitionner le tableau des mesures. Dans son exposé de 1986, Laurencelle mentionne que les désavantages du **Kappa** et des mesures d'accords «2x2», soit la domination de leurs valeurs par la fréquence des catégories, sont évités en utilisant des mesures d'accords «non pondérées» pour apprécier la "vraie compétence moyenne du codeur" (p.19). Il propose le coefficient **Pi** de Scott ou une méthode

plus simple consistant à faire la moyenne de la fidélité par catégorie, c'est-à-dire la moyenne des pourcentages d'accords pour toutes les catégories.

Une étude de Laurencelle (1983) analyse plus en profondeur le problème des accords résultant du hasard en proposant un modèle du comportement des juges dans leur tâche d'observation. Ce modèle considère la capacité des juges à bien identifier les phénomènes à observer et la possibilité qu'ils choisissent un code ou une catégorie au hasard lorsqu'ils n'y parviennent pas. Le choix de catégories au hasard est étudié selon deux contextes probabilistes, i.e. qu'il s'effectue également ou inégalement sur les catégories. Une appréciation des taux de choix réels et concomitants (lorsque deux juges ont bien identifié), d'accords relatifs (un juge identifie bien et l'autre agit au hasard) et d'accords de hasard (deux juges ont choisi au hasard) permet au chercheur d'obtenir une interprétation de la proportion d'accords entre les résultats des juges.

La conjugaison de ces deux dernières méthodes, soit la moyenne des pourcentages d'accords des catégories et l'interprétation de la proportion d'accords entre les juges, offre des solutions pertinentes aux lacunes des méthodes précédentes; le chercheur peut en effet mieux se prononcer sur la valeur de ses données lorsqu'il dispose d'une évaluation adéquate de la compétence moyenne de ses observateurs, de même que s'il a accès à une analyse qualitative des proportions de choix aléatoires.

e. Autres mesures d'accords. D'autres auteurs (Elston, Schroeder et Rojahn, 1982) se sont appliqués à concevoir une mesure d'accord qui ne soit pas

dépendante du nombre d'intervalles utilisés. Ils ont établi quatre modèles hypothétiques d'analyses s'adaptant aux types de données à traiter, selon la connaissance acquise, ou pas, de la probabilité d'incidence d'un événement dans un point temporel donné. Ces modèles assument que le taux d'erreurs d'omission est pratiquement inexistant pour avoir bien été contrôlé par l'entraînement et par des définitions d'événements validées.

Bien que les données observationnelles obtenues par l'enregistrement continu ont souvent été soumises à l'analyse de fidélité par une segmentation artificielle en intervalles de temps, des efforts sont faits pour les considérer d'une autre façon que des scores de séance. Trois études récentes discutant des problèmes d'analyse de la fidélité des données enregistrées de façon continue ont été recensées: (1) l'une propose une solution à la structuration des données observationnelles recueillies en milieu naturel par des méthodes de narration; (2) une autre utilise une extension du kappa pour obtenir une mesure de fidélité des données nominales enregistrées de façon continue, en temps réel; (3) une dernière porte sur la détection des bornes de mouvements corporels dans une tâche de micro-observation et elle présente une approche originale pour évaluer la fidélité de tels enregistrements.

L'analyse des données narratives de Scott et Hatfield (1985) implique la réduction des données à deux niveaux: la définition des unités aux descriptions données par les observateurs, suivie de la catégorisation par les analystes. Il en découle deux niveaux d'analyse de la fidélité, soit l'accord entre observateurs sur l'épisode comportemental ou l'unité d'activité, et l'accord

entre analystes sur le codage des épisodes ou des unités. Scott et Hatfield (1985) ont formulé un modèle d'analyse tenant compte de la durée (en plus de la fréquence) et du chevauchement des événements décrits. Ils comparent la durée d'accord de chaque unité à l'accord total (% d'accords effectifs), ou à la durée moyenne (estimé de précision) de tous les épisodes ou unités codés (vs durée de la séance d'observation). Leur formule d'accord proportionnel comporte un facteur d'ajustement du temps pour chaque unité. En conclusion, ces auteurs stipulent qu'avec des unités d'observation de longueurs inégales, cet ajustement de la formule d'accords en fonction de la durée contribue à augmenter la validité écologique[1] de l'estimé. De plus, ils soulignent que des taux plus élevés d'accords sont anticipés avec des épisodes comportementaux relativement plus longs en comparaison de l'étendue de la séance d'observation. Dans les cas où les résultats sont composés principalement de ces unités, ils suggèrent qu'une correction pour les accords imputables au hasard soit apportée.

Une autre étude (Conger, 1985) s'appliquant aux observations de comportements ou d'événements enregistrés de façon continue, disons aux échelles nominales «continues», utilise le coefficient de fidélité 'K' dans le but de procurer une mesure de fidélité non arbitraire . L'auteur propose que les événements soient enregistrés en temps réel et que les indices de fidélité soient basés sur l'accord «point par point» ("moment by moment agreement").

---

[1]Cette appellation de la validité fait référence aux études d'observation écologique introduites par Barker et Wright (1955; voir Cone 1982) comme une méthodologie particulière en éthologie sociale.

La statistique **K** est une extension du kappa ('**k**') de Cohen, lequel, à sa différence, est employé avec des scores de fréquences provenant d'intervalles discrets divisés arbitrairement. Donc '**K**' se base sur les durées d'événements concordants alors que '**k**' utilise les fréquences d'événements concordants. Les données provenant des comportements codifiés de façon continue se définissent comme les intervalles de temps qu'on a attribués aux états particuliers (présents/absents, etc.) observés chez l'objet-cible. L'accord inter-observateurs est obtenu dans les intervalles de temps où deux observateurs assignent le même état; il s'agit d'une même catégorie avec des durées communes. Toutefois, Conger (1985) explique que l'usage du '**K**' implique des problèmes d'ordre théorique et pratique dus à la nécessité d'identifier un nombre potentiel d'observations indépendantes, et dus à la sensibilité du système d'encodage. L'indépendance des événements n'est idéalement limitée que par la sensibilité de la réaction de l'instrument d'observation (dans certains cas, le codeur lui-même et dans d'autres, le codeur et l'équipement d'encodage). Ainsi, lorsque les observateurs codent sur le vif, le temps de préparation pour une observation subséquente influence le nombre potentiel d'observations indépendantes. Par exemple, si on calcule .04 seconde comme période de recouvrement d'un micro-ordinateur avec lequel les informations sont emmagasinées directement, on divisera le temps total d'une séance par ce nombre pour connaître la quantité potentielle d'observations distinctes; et cela même si le temps de détection d'un événement par un observateur est inférieur au temps requis par l'ordinateur pour accepter de nouvelles données (cf. Conger, p. 867). En dehors des circonstances observationnelles idéales, Conger explique

qu'il est plus probable que le nombre d'observations indépendantes soit davantage limité par la capacité d'attention et de réaction des observateurs que par le système d'encodage; c'est pourquoi il ne recommande pas de prendre le nombre total d'événements observés comme la base aux tests de signification ou aux intervalles de confiance. Donc, en l'absence de l'information concernant le nombre total d'observations indépendantes, l'auteur recommande que 'K' soit considéré seulement comme une statistique descriptive qui exprime en termes approximatifs la fidélité de l'échelle nominale continue.

L'originalité de la dernière étude par McDowall (1978) réside dans la non-fixité de l'unité d'accord. La compétence des observateurs dans leur détection des bornes de mouvements à partir de séquences filmées fait l'objet de la fidélité. Cette étude se caractérise par une microanalyse des mouvements corporels, l'attention étant portée particulièrement sur les débuts, les fins et les changements de mouvements dans 15 parties du corps. Chaque image de la séquence filmée représente l'unité discrète sur laquelle se base l'accord des enregistrements; l'auteur a constitué 15 séquences (soit une pour chaque partie du corps) de 50 images. L'accord y est défini comme la coïncidence de bornes repérées au même endroit; l'accent étant placé sur la discrimination des modifications dans le cours du mouvement, peu d'importance est attribuée à l'habileté à classifier correctement les bornes. Cependant, l'unité d'accord a aussi été extensionnée à plus d'une image (jusqu'à six) dans le but de déterminer un intervalle judicieux compensant pour les différences entre observateurs. Les résultats montrent qu'une probabilité d'accord

significative est difficilement atteignable avec un intervalle d'une image et que l'extension de l'unité d'accord offre des indices de fidélité plus acceptables en termes de grandeur et de signification. McDowall considère qu'un intervalle de trois images est suffisant pour retrouver des probabilités d'accords significatives. Aussi, il reconnaît qu'un accord basé sur des unités trop étendues diminue la sensibilité des mesures de fidélité. Toutefois, rappelons qu'il s'agit ici de «fidélité microcinétique», où l'unité de base, une image, correspond à 1/24 de seconde; donc, même avec un intervalle de 6 images (1/4 de seconde) comme unité d'accord, nous sommes encore beaucoup plus précis que dans l'observation aux 10 secondes d'intervalles.

Les travaux de Laurencelle (1981) montrent des analyses mathématiques plus poussées dans l'utilisation des propriétés métriques de la fidélité d'accords. Cet auteur a formulé plusieurs modèles de fidélité se basant sur trois niveaux de mesure (global=pour tout un processus, par item=pour chaque observation, et par catégorie=pour chaque cote). Il y développe cinq modèles de fidélisation des données à partir du simple pourcentage d'accords inter-juges:
(1) "La proportion d'accords comme coefficient de fidélité". Avec les systèmes de cotation utilisant des étiquettes exclusives pour décrire l'objet d'observation, l'auteur présente la formule de base du «pourcentage d'accords»: **PA= #accords/#événements codés**. C'est un estimé de la fidélité globale qui donne la proportion des accords sur les accords possibles. Une autre formule est présentée pour les sytèmes à cotes multiples:

(F10)

$$PA = \frac{\Sigma \text{ cotes identiques entre (2) observateurs pour chaque item}}{\Sigma \text{ (cotes obs.}_1 \text{ x cotes obs.}_2)^{1/2} \text{ pour chaque item}}$$

Un cas particulier est aussi étudié pour "estimer la cohérence des cotes de deux à plusieurs juges pour un item donné".

(2) "La fidélité par catégorie avec critère de vérité". Un modèle probabiliste d'appréciation de la compétence d'un observateur par rapport à une liste de cotes "vraies" est développé pour les sytèmes à cote simple. C'est une mesure du degré de réalité qu'un juge présente dans ses données.

(3) "La fidélité de consensus pour des cotes multiples". C'est une autre construction probabiliste des attributions aléatoires par deux juges pour estimer l'accord global, i.e. sur tous les items.

(4) "La fidélité par item avec critère de vérité et catégorie «poubelle»". L'originalité ici est de fournir à l'observateur une «pseudo-catégorie» pour les cas où il hésite dans son choix de catégories. L'inclusion d'une catégorie «poubelle» dans un système d'observation a un double avantage: développer l'assurance de l'observateur dans son choix des catégories véritables et, ainsi, enregistrer les phénomènes plus à l'image des catégories préétablies; permettre au concepteur de vérifier ses définitions par l'examen de ce que les juges ont écarté à la «poubelle». L'indice de fidélité fourni par ce modèle représente réellement l'habileté d'un observateur à choisir une bonne catégorie.

(5) "Un modèle d'interprétation de la fidélité de consensus". L'auteur a développé des formules pour transformer un coefficient de fidélité par critère de consensus en un estimé de la fidélité par critère de vérité, de façon à faciliter son interprétation.

House, House et Campbell (1981) apportent leurs conclusions sur les différentes statistiques de l'accord entre observateurs à partir des comparaisons qu'ils ont effectuées entre plusieurs indices calculés sur les mêmes scores provenant d'une multitude de résultats possibles. Leurs comparaisons considèrent quatre types de variation dans les mesures d'accord: (1) la fréquence d'événements présents (A+B et A+C); (2) la fréquence d'erreurs (quantité de désaccords entre observateurs, soit B+C); (3) la distribution des erreurs selon que les cases B et C diffèrent ou sont identiques (B>C ou B<C ou B=C); (4) la fréquence d'événements absents (quantité d'occasions où un comportement n'est pas enregistré, soit B+D et C+D). Les résultats de ces comparaisons montrent entre autres que les valeurs obtenues avec le pourcentage d'accords «2x2», le **kappa** et le **phi** sont sous deux influences: le taux global d'occurrences du comportement-cible et la répartition égale entre accords d'événements présents et d'événements absents pour un nombre donné de désaccords. Ils déclarent que le fait de corriger pour les accords imputables au hasard n'a d'autre effet que de rendre plus difficile l'obtention de hauts niveaux d'accords avec des fréquences de comportements élevées ou faibles. Ils soulèvent un autre aspect inquiétant pour ces mêmes mesures (K, Ø, %A): les indices prennent de plus grandes valeurs lorsque la balance des désaccords

entre observateurs est très disproportionnée (e.g.: B=10 et C=0). Cette tendance à la hausse des indices est préoccupante puisque les chercheurs ne peuvent ainsi pas relever le biais sérieux présent dans le codage et reflété par ce type d'erreurs. Enfin, ils observent que certaines mesures sont symétriques par rapport aux accords d'événements présents vs événements absents: la valeur de l'indice sera la même pour A=20, B=5, C=5, D=70 que pour A=70, B=5, C=5, D=20 avec les mesures du **kappa**, du **phi**, ou du **pourcentage d'accords total**, entre autres. Cependant, les mesures comme le **lambda**, le **pourcentage d'événements présents** et le **pourcentage d'événements absents** sont asymétriques puisque certaines cases sont ignorées. En bref, ils mentionnent qu'il n'y a pas de meilleures mesures et qu'elles doivent toutes être utilisées avec circonspection selon différentes considérations. Avec les tableaux «2x2», ils suggèrent une option qu'ils jugent plus intéressante: rapporter les moyennes des cases A, B, C et D parmi les observations (formule **20**, matrice d'accords moyens). Ils soutiennent que les mesures d'observation directe sont solidement ancrées dans les habitudes de recherche et que des solutions aux problèmes de leur appréciation doivent être trouvées.

Les nombreuses critiques adressées à l'endroit des formules de pourcentage d'accords sont un reflet de l'inaptitude des techniques existantes à se rattacher à une conceptualisation générale de la fidélité. Des suggestions plus «contemporaines» ont été présentées dans les travaux de Laurencelle (1981,83,86) et elles touchent plus judicieusement aux problèmes actuels rencontrés dans la procédure de vérification des données observationnelles.

Cet auteur prétend que deux mesures suffisent à satisfaire les questions de fiabilité des codeurs et des données, ce qu'il appelle le "contenu de vérité du processus observationnel": (1) l'utilisation des mesures d'accords non pondérées (e.g.: coefficient pi de Scott, ou le pourcentage d'accords moyen pour toutes les catégories) pour apprécier la vraie «compétence moyenne» du codeur est recommandée plutôt que les indices de fidélité pondérés (parce que la valeur de ces dernières mesures est infléchie par les catégories (ou scores) plus fréquentes; (2) le calcul de la part imputable au hasard dans le pourcentage d'accords; Laurencelle a conçu un modèle à priori de la réponse de hasard pour ensuite obtenir une mesure plus «raffinée» de la fidélité, i.e. avec les contributions aléatoires en moins (Laurencelle, 1986). Un exemple d'opérationnalisation d'une méthode de fidélisation des données développée par Laurencelle, se retrouve dans le rapport méthodologique d'une étude séquentielle des comportements d'enfants en interaction dyadique. La valeur statistique du pourcentage d'accords et son degré de signification y sont appréciés par deux méthodes: (1) la fidélité des codes simples, comportant trois analyses statistiques différentes (% d'accords, X^2 reflétant la probabilité empirique des codes produits, et la signification statistique du pourcentage d'accords observé après permutations aléatoires); (2) la fidélité des codes regroupés, conçue pour pallier à la catégorisation restrictive en cotes simples, utilise les mêmes statistiques de la première méthode. Dans cette étude, Laurencelle fournit même une comparaison avec la fidélité «présence/absence» en créant un programme statistique capable d'analyser les mêmes données avec cette autre approche plus connue et plus traditionnelle.

Après cette longue discussion sur les différentes méthodes de pourcentages d'accords pour apprécier la fidélité des données nominales, nous complétons ce tableau avec les méthodes corrélationnelles qui sont applicables aux données quantitatives.

2. Méthodes corrélationnelles. Les mesures corrélationnelles permettent aussi d'estimer la fidélité par critère de consensus entre deux ou plusieurs observateurs et elles auraient en plus l'avantage de comporter un certain contrôle pour l'accord attribuable au hasard. House, House et Campbell (1981) expliquent que ce contrôle s'exerce dans la dévaluation de la concordance atteinte à des taux d'occurrences élevés ou faibles, par opposition aux taux modérés se situant autour d'une fréquence d'occurrence de 50%.

Les mesures corrélationnelles trouvent leur utilité dans la comparaison des totaux sur les intervalles de temps ou sur les individus. Elles sont applicables aux données de fréquence ou de durée des catégories de comportement mais pas aux occurrences individuelles. La corrélation est établie entre deux listes de scores ordonnés sur un même sujet ou entre les scores de deux observateurs sur un nombre d'intervalles de temps différents pour un sujet. Dans leur étude de 1979, Caro et ses collègues interprètent le coefficient de corrélation entre observateurs comme une mesure de l'habileté des observateurs à discriminer entre les individus, tenant compte à la fois de l'erreur due aux observateurs et des différences entre le comportement des individus dans l'échantillon observé. Cette approche d'évaluation de la fidélité inter-juges sert spécifiquement aux études sur les différences individuelles

car elle permet de vérifier si les différences entre individus sont plus grandes que les différences entre observateurs.

Dans son tableau des principaux indices d'estimation de la fidélité, Beaugrand (1982) identifie la corrélation de rangs de Spearman et le coefficient de concordance de Kendall (W) comme les indices associés aux systèmes avec mesures ordinales provenant de deux observateurs. Berk (1979) mentionne que la corrélation de Pearson, utilisée pour estimer la fidélité par séance, est plus souvent employée que la corrélation de rangs de Spearman; en référence à Hartmann (1977), la corrélation de Pearson sera détaillée plus loin. Le 'W' de Kendall s'applique aux cas avec plus de deux observateurs; le 'W' est une statistique descriptive et son intérêt est d'obtenir le degré auquel les rangs attribués par plusieurs observateurs s'approchent de l'accord maximum. Les scores quantitatifs des catégories de comportement sont, au préalable, réduits à des rangs, classes ou ordres. Les mesures sur échelles à intervalles sont évaluées avec la corrélation de Pearson (2 observateurs), ou la corrélation intraclasse (plus de 2 observateurs). Le **A** de Robinson est aussi indiqué pour les échelles à intervalles. Il s'applique particulièrement aux échelles d'évaluation ("rating scales"). Le **phi (ø)** est une simplification de la corrélation de Pearson appliquée aux données dichotomiques (présence/absence). Cette corrélation s'applique aux présences-absences d'un seul comportement, ou catégorie, et il peut être calculé à partir des valeurs aux cases du tableau «2x2»:

(F11)

$$\phi = \frac{(AD-BC)}{[(A+B)(C+D)(A+C)(B+D)]^{1/2}}$$

Hartmann (1977) démontre que les valeurs des statistiques **kappa** et **phi** sont presque identiques lorsque le taux d'événements codés présents est approximativement égal entre les deux observateurs. Lorsque le ϕ équivaut au **kappa**, il est interprété comme une statistique de **pourcentage d'accords corrigé**. Pour Hollenbeck (1978), le coefficient de corrélation de Pearson détient plusieurs avantages sur le pourcentage d'accords par ses propriétés métriques connues et par sa relation avec plusieurs autres techniques mathématiques et statistiques. Toutefois, son utilité est restreinte aux données de type ordinal. Les mêmes restrictions que pour le Kappa, formulées par Towstopiat (1984), s'appliquent au Phi, avec un désavantage en plus pour ce dernier: la corrélation n'est pas définie s'il n'existe pas de variabilité entre les réponses des observateurs.

Hartmann (1977) représente le coefficient de fidélité inter-observateurs des données de séance ("session scores") par le symbole r_{kk}, alors qu'il utilise le symbole r_{xx} pour le coefficient de fidélité des données d'événements séparés ("trial scores"). Cette désignation de la corrélation de Pearson ne doit toutefois pas être confondue avec celle de la théorie classique des tests; il s'agit simplement pour Hartmann de distinguer l'index corrélationnel des mesures de séance de celui des mesures d'événements séparés. Ainsi, Hartmann

explique que le r_{kk} est calculé sur les scores pairés provenant des séances communes (ou des données quantitatives) observées par deux juges. Le coefficient r_{kk} prend typiquement des valeurs entre '0' et +1.00[1], indiquant la plus ou moins grande présence de relation entre les évaluations des deux observateurs. Hartmann (1977) donne les interprétations mathématiques du r_{kk}: il égale la proportion de variance du score total qui n'est pas due à l'erreur et le degré d'association linéaire entre les données des deux observateurs. Le r_{kk}^2 égale la proportion de variance des scores d'un observateur qui est prévisible en connaissant les scores de l'autre observateur. Les principaux avantages de cette méthode corrélationnelle sont énumérés par Hartmann comme suit: la possibilité de produire un intervalle de confiance indiquant la plus petite différence entre les scores de séance qui puisse être interprétée de façon significative; avec certaines restrictions, la description précise du degré de dépendance linéaire ou de corrélation entre les évaluations des observateurs; les propriétés du r_{kk} sont bien connues et d'autres tests statistiques peuvent être appliqués. Le désavantage majeur dont Hartmann fait mention se retrouve lorsque la variabilité des scores est zéro pour un ou pour les deux observateurs, cas auquel la valeur de r_{kk} est indéterminée. Quelques-unes des autres limites ont trait à l'interprétation du r_{kk} dans des situations particulières (e.g.: erreurs correlées entre les observateurs, etc.) et à l'influence de l'étendue des scores sur la valeur de r_{kk}.

---

[1] L'étendue possible du r_{kk} est - 1,00 à + 1,00 mais des coefficients de fidélité négatifs sont théoriquement impossibles et empiriquement rares.

Un rapport peut être établi entre la fidélité des mesures d'événements séparés ("trial reliability") et la fidélité des mesures de séance ("session reliability"): son explication est élaborée par Hartmann (1977). Hartmann parle d'un lien «formel et mathématiquement précis» dans le cas où la corrélation de Pearson a été calculée sur les scores d'événements séparés; cette mesure corrélationnelle fournit un estimé de base pour la fidélité des scores de séance, ces derniers étant des composés formés des scores d'événements séparés. Le coefficient r_{kk} des scores de séance sera habituellement plus élevé si le coefficient r_{xx} a une valeur moyenne égale à 0.6. Cependant, cette relation entre les scores d'événements séparés et les scores de séance ne peut pas être établie avec les statistiques d'accord. Néanmoins, une recommandation de Hartmann stipule que les chercheurs devraient présenter leurs statistiques de fidélité de façon à permettre le transfert d'une expression statistique à une autre. Les investigateurs auraient alors le loisir de calculer la statistique qui les intéresse.

Selon Johnson et Bolstad (1973), les méthodes corrélationnelles sont utiles dans les situations suivantes: (1) lorsqu'on ne peut être sûr que les mêmes comportements sont conjointement observés au même temps; (2) lorsqu'un niveau élevé d'accords, dus à la chance uniquement, est probable; ou (3) lorsqu'on ne dispose que d'un échantillon limité de données d'accords inter-observateurs par rapport à l'ensemble des données d'observation recueillies. Les dangers, selon ces mêmes auteurs, résident dans l'interprétation des coefficients. En effet, il est possible d'obtenir des coefficients de corrélation

élevés lorsqu'un observateur, par rapport à un autre, donne une surestimation constante des fréquences comportementales. Il est aussi plus probable d'obtenir des valeurs élevées de corrélation avec l'augmentation de l'étendue des scores de la variable dépendante en dépit de l'écart des observateurs dans le nombre d'occurrences des comportements observés. Hartmann (1977) maintient que les mesures corrélationnelles sont préférables aux mesures de pourcentages sur la base de leurs propriétés mathématiques (correction pour l'accord dû à la chance) et, en plus, de leur adaptabilité à théorie de la généralisabilité. L'étude de généralisabilité des données, conduite avec les mesures de corrélations intraclasses, est recommandée par bon nombre d'auteurs (Berk, 1979; Cone, 1977, 82; Foster et Cone, 1980; Hartmann, 1982; Mitchell, 1979) parce qu'elle sert à évaluer les contributions des diverses sources de variation présentes dans les données. Les travaux de Berk (1979) offrent une bonne critique des intérêts et avantages de calculer les coefficients de généralisabilité. Il mentionne entre autres que la théorie se prête particulièrement à la «multidimensionnalité du comportement adaptatif» et à la «multiplicité des facteurs influençant l'estimation et l'interprétation de la fidélité inter-juges» (p. 464). Parmi la liste des avantages énumérés par l'auteur, il est intéressant de retenir que l'analyse de généralisabilité procure des estimations non biaisées de la fidélité inter-juges pour une seule observation, et pour des ensembles d'observations, et qu'elle peut être appliquée à une variété de systèmes d'observation dits «quantitatifs» et «de catégorie».

Dans son article de 1979, Mitchell déclare qu'une étude de généralisabilité. est préférable au calcul d'un coefficient de fidélité. Ses raisons sont en premier que les coefficients de généralisabilité apportent plus d'information utile à propos des sources de variabilité dans un ensemble de données; et ensuite, que le plan conceptuel d'une étude de généralisabilité, par ses différentes facettes, convient à d'autres sortes d'analyses pour la plupart des études observationnelles (e.g.: coefficient de généralisabilité se rapportant aux observateurs, coefficients pour les sujets individuels, etc.), en plus de correspondre au plan global de la recherche. En outre, Mitchell mentionne qu'une étude de généralisabilité combine deux intérêts importants pour le chercheur sur le comportement humain: les différences individuelles entre les sujets et l'influence d'autres facteurs (habituellement environnementaux) sur le comportement.

Enfin, Hartmann (1982) formule trois ordres de recommandations ayant trait au contenu de l'information à publier sur les analyses de fidélité des scores: (1) les types de fidélité à rapporter devraient inclure la précision et la consistance inter-juges en plus de la fidélité de séance; (2) les sources de données servant aux calculs de fidélité devraient être constituées de vérifications périodiques et dissimulées sur différents sujets et dans différentes conditions; et (3), des classes de scores devraient être utilisées dans le rapport de fidélité, i.e. chaque variable-cible de l'analyse comportementale, individuelle ou composée selon le cas ferait l'objet d'une classe séparée d'analyse.

D'un point de vue général, les méthodes d'évaluation de la fidélité se différencient par le type de données qu'elles traitent: les méthodes d'accords inter-juges sont appliquées aux données catégorielles et les méthodes corrélationnelles sont surtout utilisées avec les données quantitatives[1]. On reproche aux premières de ne pas avoir de lien avec la conception classique de fidélité et on utilise les secondes en assumant que les cotes des observateurs sont des mesures comparables aux scores individuels, ou résultats, à un test. Alors on juge que les coefficients d'accords inter-juges ne sont pas des indices de la fidélité inter-juges et on en conclut que l'information qu'ils apportent est accessoire. Par conséquent, parce qu'on persiste à faire fi de la nature exclusive des données catégorielles provenant des techniques d'observation directe, on éloigne l'opportunité de considérer une nouvelle conception théorique pour cette méthode particulière d'acquisition des connaissances. Et même si les techniques d'observation sont devenues plus systématiques, les données n'en obtiennent pas plus un statut scientifique parce que l'évaluation de leur véracité repose toujours sur des méthodes contestées et parfois, inappropriées. Cette problématique sera commentée davantage au dernier chapitre et elle constituera le motif de présentation d'une nouvelle conception théorique de la fiabilité des données en observation directe. Auparavant, cette section est complétée par une introduction succinte aux modèles d'interprétation des indices de fidélité ainsi que par une brève conclusion touchant au problème de la validité des données.

---

[1] Les intercorrélations et les corrélations intraclasses peuvent aussi être calculées avec les données de catégorie pourvues d'un code numérique (Berk, 1979).

Quelques modèles d'interprétation des indices de fidélité

Les modèles d'interprétation de la valeur d'un indice de fidélité mentionnés dans l'article de Hartmann (1982) sont développés ci-après:

A. Les échelles de probabilité conditionnelle

Les échelles de probabilité conditionnelle sont illustrées par l'indice d'accords de Dice (S_D); $\mathbf{S_D}$ est la probabilité conditionnelle qu'un second observateur ait codé la présence d'un événement en prenant pour acquis que le premier observateur en a codé la présence aussi. Cette perspective a été davantage examinée par Bergan (1980, voir Hartmann, 1982) et par Towstopiat (1984), qui avec Bergan, a développé des mesures d'accords univariées et multivariées basées sur les modèles et les estimations de quasi-équiprobabilité et de quasi-indépendance. L'analyse des modèles univariés est appliquée dans le cas où deux ou plusieurs observateurs codent plusieurs comportements à un point dans le temps. Dans les situations où des observations sont faites parmi différents points dans le temps et à travers différentes conditions de traitement, les techniques d'analyse multivariée sont employées. Towstopiat (1984) prétend que ces modèles améliorent la précision des mesures de trois façons: (1) en donnant un coefficient d'accords basé sur la probabilité avec une signification directement interprétable; (2) en corrigeant pour la proportion d'accords imputables à la chance; et (3) en permettant la séparation des estimations d'accords et de désaccords à l'intérieur des modèles.

Un autre modèle probabiliste, postulé par Laurencelle (1983) et mentionné précédemment, explore l'aspect comportemental des juges en se basant sur des modèles de choix aléatoires parmi les catégories (à nombres variés) d'un répertoire préétabli. Dans une autre étude de Laurencelle (1981), un coefficient d'accords probabiliste est élaboré, mais pour des systèmes à cotes multiples[1].

B. Les échelles de proportion d'accords

Les échelles de proportion d'accords sont représentées par le simple pourcentage d'accords et les autres statistiques d'accords qui en dérivent. La mise en garde faite par Hartmann tient à l'interprétation des valeurs '0' sur ces échelles; la signification du zéro peut correspondre à l'absence d'accords dans le cas des statistiques d'accords simples, ou elle peut indiquer un niveau d'accord égal à celui anticipé par la chance avec les statistiques corrigeant pour les accords imputables à la chance (pi de Scott, kappa de Cohen). Laurencelle (1981) a développé un modèle d'interprétation de la fidélité de consensus, lequel est comparable au modèle de la théorie classique mettant en relation «variance observée» et «variance vraie». Dans le modèle de Laurencelle, on obtient une estimation de la «fidélité de vérité moyenne» des juges à partir de la «fidélité de consensus», c'est-à-dire le pourcentage d'accords.

---

[1] Un juge peut choisir plusieurs cotes pour le même événement observé, contrairement aux systèmes à cote simple, où chaque catégorie décrit un seul objet d'observation.

C. Les échelles de proportion de variances

Les échelles de proportion de variances sont utilisées dans les statistiques de fidélité corrélationnelles, telles que les corrélations intra-classes dans l'approche de la théorie de la généralisabilité, et elles décrivent la fidélité comme le rapport de la vraie variance à la variance obtenue. La formule équivalente de la théorie classique de la fidélité est la proportion de variance du score vrai contre la variance du score d'erreur ajoutée au score vrai (**V/V+E**), mesurée à travers la corrélation de Pearson.

Validité des données observationnelles

La littérature à propos des recherches comportant des procédures d'observation directe fait peu état du problème de la validité des mesures ainsi obtenues. Bien que le concept de validité soit bien développé dans le cadre de la théorie classique des tests, il est encore loin d'être appliqué aux données observationnelles. Laurencelle (1986) mentionne l'inexistence d'études approfondies en matière de validité interne des mesures catégorielles et il en conclut que le problème de la «véracité» des observations a été plutôt approché sous l'angle de la fidélité, soit par la mesure de concordance entre deux listes d'observations indépendantes.

Cone (1982) replace le concept classique de validité dans le contexte de l'observation directe et il réconcilie les différents types de validité (de contenu, de construit, de critère, écologique, convergente, discriminante, de traitement) aux systèmes d'observation directe. Pour d'autres auteurs (Berk,

1979; Cone, 1977; Foster et Cone, 1980; Hartmann, 1982, Mitchell, 1979), l'application de la théorie de la généralisabilité aux données observationnelles constitue une bonne évaluation de la validité des données. L'analyse des différents «univers», ou contextes de généralisation, est interprétée en termes de «validité écologique»[1]: la «vraie» quantité, ou propriété dimensionnelle du «vrai» comportement est observée dans le «bon» (vrai) environnement (Foster et Cone, 1980).

En résumé, les auteurs s'accordent pour définir conceptuellement la validité des procédures observationnelles comme «l'objectivité avec laquelle les comportements sélectionnés dans un système observationnel reflètent et échantillonnent le problème concerné». Sur le plan pratique, les méthodes d'évaluation tardent à apparaître. Un seul procédé d'étude de validité non traditionnelle a pu être recensé[2]; la représentativité des événements codés y est abordée par un principe de reconstruction des observations; les images obtenues à partir des reconstructions sont à nouveau codées par un observateur différent pour être ensuite comparées aux observations initiales. Toutefois, cette approche comporte des limites dont, entre autres, la possibilité de conserver le biais initial introduit par un premier juge, l'objectivité des données n'en n'étant pas mieux évaluée. La facilité d'obtention d'un seuil

---

[1]La validité écologique d'une recherche tient de la capacité de ses concepts et de ses techniques d'analyses à bien représenter le système social étudié (Barker, 1965 et Bronfenbrenner, 1977; voir Trudel et Strayer, 1986).

[2]Le procédé de reconstruction d'un protocole d'observation pour un sujet donné a été initié par Frey et Pool (1976) et il a été repris dans les travaux de Déziel (1985) comme méthode de vérification de la validité des mesures à bas niveau d'abstraction. Le lecteur peut se référer à cette dernière auteure pour de plus amples détails.

d'accords élevé avec la méthode de reconstruction démontre qu'une certaine censure des phénomènes observés initialement a pu être conservée ou reproduite dans les enregistrements utilisés. De plus, la difficulté de codification est allégée dans le cas des observations sur photos; la concordance devient ainsi augmentée.

Le chapitre suivant présentera une nouvelle conceptualistion du processus observationnel ainsi que divers modèles d'opérationnalisation de la fidélité instrumentale pour des données multidimensionnelles et continues.

Chapitre III

Un modèle d'observation systématique et une nouvelle conception de la fidélisation des données

Dans le premier chapitre, nombre de systèmes d'enregistrement du comportement ont été présentés. Chacun à sa façon, avec plus ou moins de précision, de complexité, d'exhaustivité et d'objectivité, parvient à décrire un ou plusieurs aspects du phénomène comportemental. Au deuxième chapitre, une énumération des diverses méthodes statistiques pour évaluer la qualité ou la fiabilité des observations codées par des juges, sous le schème conceptuel du système utilisé, nous fait voir les faits suivants: (1) Il n'existe pas de consensus quant à la conception théorique du processus de vérification des données; certains recherchent des solutions nouvelles et d'autres tentent d'adapter les anciennes théories aux procédures actuelles. (2) L'état de la recherche en ce domaine est à l'époque des essais empiriques; une conséquence en est l'existence d'une multitude de propositions différentes et souvent incompatibles. (3) Les systèmes d'observation conceptuellement bien articulés produisent des données pour lesquelles la statistique n'a pas trouvé des techniques d'analyse adaptées; le résultat est que pour rendre compte de la valeur de leurs données observationnelles, les chercheurs doivent restreindre leur plan d'observation à un minimum de dimensions. La reproduction du phénomène comportemental par l'observation demeure ainsi encore segmentée et largement réduite.

Toutefois, la carence de moyens appropriés et l'incertitude des méthodes suscitent chez plusieurs l'intérêt à l'investigation. Particulièrement dans le domaine de recherche sur le comportement non verbal[1], bon nombre de chercheurs ont ébauché des méthodes de cueillette des données plus sophistiquées. Les principales études récentes seront les seules retenues, parmi lesquelles deux tendances se dessinent quant à l'objectif de décrire le comportement non verbal[2], soit la microanalyse et la macroanalyse. La première a pour but de différencier les aspects minuscules ou fins du comportement; par exemple, l'étude cinétique du mouvement, ou la description gestétique d'un geste de la main; la seconde identifie le phénomène à grande échelle comme dans les descriptions macroscopiques. De plus, deux approches définissent les stratégies adoptées par les chercheurs: il s'agit de l'approche structurale dans laquelle on vise à établir un répertoire global informant sur l'organisation et la hiérarchisation des processus comportementaux (e.g.: un système de catégories décrivant les caractéristiques spatio-temporelles de mouvements particuliers); l'approche fonctionnelle va plus loin que la description des patrons comportementaux; elle utilise les détails sur la nature des signes non verbaux pour définir le critère fonctionnel des patrons

---

[1] Le lecteur est ici référé au mémoire de maîtrise de Déziel (1985), dans lequel un vaste recensement de systèmes est fourni conjointement aux orientations spécifiques du secteur de recherche de la communication non verbale.

[2] Le terme "non verbal" est pris dans son sens générique, i.e. sans sélection sur les multiples aspects qui le concernent (e.g.: postures, gestes, contact des yeux, distance et proximité physiques, mouvement corporel, etc.).

particuliers de comportement (e.g.: comportements d'adaptation, de survie, etc.).

Dans Cosnier (1984), on retrouve une mention des principales études se caractérisant par ces quatre aspects: l'étude des façons de se mouvoir et d'utiliser son corps, i.e. la «kinésique» de Birdwhistell, et l'analyse séquentielle des événements verbo-moteurs par Condon, nous font découvrir des exemples de microanalyses structurales. L'étude des mimiques faciales par Ekman et Friesen, ainsi que l'analyse des signes non verbaux dans la communication par Scherer sont des prototypes de microanalyses fonctionnelles. Les méthodes macroanalytiques pour leur part, se sont inspirées directement de l'éthologie animale. Les travaux de Blurton Jones, de McGrew et de Montagner en sont des exemples (Cosnier, 1984).

Première partie: Présentation du système d'observation Somac

Des modèles d'observation systématique

Les propos qui précèdent nous permettent de constater qu'un chercheur ayant l'ambition de concevoir un système de représentation et de transcription du comportement non verbal avec des critères de spécificité, de qualité et de quantité doit s'astreindre à un plan d'observation rigoureux. Par exemple, il appliquera les règles suivantes dans la définition de ses catégories:

description moléculaire du comportement, minimum d'inférence, bas niveau d'abstraction, non-sélectivité, inclusion de tous les comportements. Des analyses comportementales avec un niveau de spécificité aussi élevé ont déjà été entreprises malgré les contraintes extrêmes (coûts astronomiques, disponibilité accrue du chercheur, qualités d'attention exigeantes pour l'observateur, matériel d'enregistrement et de codage sophistiqué) qu'elles imposent. Seulement deux systèmes de ce genre seront décrits pour leurs similitudes avec le système faisant l'objet de la présente recherche.

Il est question du «Berner System» de Frey et Pool (1976) et de la grille d'observation des événements verbo-moteurs de Condon (1979). Le système de Frey et Pool sert à transcrire le mouvement corporel à partir d'un principe de notations en séquences temporelles; des positions spatio-temporelles sont codées à des intervalles d'une demi-seconde. Ce système utilise des enregistrements magnétoscopiques du matériel à observer. Le mouvement est donc déduit du protocole inscrit à travers les changements de position dans le temps. L'amplitude du mouvement est prédéterminée par la définition d'unités spatiales en référence à la physionomie anatomique du sujet (e.g.: main à la hauteur de la tête). Sur le plan métrique, on rencontre ici le problème d'unités qui ne sont pas égales, contrairement aux items d'un test d'aptitude par exemple, où chaque réponse vaut un point. Cependant, l'aspect temporel des données est facilement déduit du protocole par l'identification des moments de début et de fin de mouvement; la comparaison des cotes des observateurs est donc facilitée puisque ce système crée artificiellement des intervalles de

temps; la complexité réside dans la multiplicité de cotes attribuables au même moment (positions différentes de tous les segments corporels observables) et l'accord inter-juges comporte plusieurs dimensions. Ainsi des juges peuvent avoir un accord total sur des positions de certains segments, avec une concordance plus ou moins parfaite dans le temps; ou, pour un intervalle donné, des juges peuvent s'entendre sur une partie seulement de toutes les positions identifiables. Il est donc moins probable d'obtenir un accord global parfait. Les propos de Frey et Pool, tirés de leur rapport de recherche de 1976, peuvent indiquer la pertinence du problème posé ici:

> "...le chercheur doit toujours enregistrer ou mesurer les comportements par l'intermédiaire d'une certaine forme de protocole, et c'est en définitive la précision, fidélité et «puissance de résolution» du protocole qui détermine la qualité des données utilisées dans ce genre de recherche. Alors le problème majeur dans ce domaine en est un de mesure et les difficultés réelles impliquées sont très grandes."

L'instrument de Condon se caractérise aussi par la multidimensionnalité microanalytique et comporte un codage en séquences temporelles de 1/30 de seconde. Cependant, la méthode de segmentation de l'objet à coder est différente de celle de Frey et Pool: au lieu de noter une série de positions statiques, Condon a conçu une méthode d'observation du mouvement basée sur le repérage du changement de direction et de vitesse d'une partie du corps. Les unités spatiales du corps y sont repérées au niveau des articulations par la reconnaissance anatomique de leurs modes de déplacement, soit l'extension, la

flexion, l'adduction, l'abduction, la rotation, etc. Par la transcription des changements dans les différentes parties du corps, Condon obtient des configurations complexes des parties du corps lui servant dans son étude structurale, i.e. l'analyse organisationnelle des relations de changement entre les parties du corps[1]. L'effort de Condon consistait donc à concevoir une méthode de segmentation de la chaîne comportementale pour découvrir l'information sur sa structure, i.e. comment les éléments apparemment discontinus s'organisent en processus continus (Condon, 1984). Mais qu'en est-il de l'analyse de la valeur métrologique d'un tel instrument? Condon mentionne seulement qu'il s'agit d'une étude éthologique. Il est évident que plusieurs difficultés méthodologiques sont encourues par un projet d'analyse statistique de la fiabilité des données enregistrées par un tel procédé. Un premier problème à surmonter réside dans la définition de l'unité d'accord: comment déterminer un accord basé sur une unité d'observation multidimensionnelle?

A. Particularités de l'instrument Somac

Le système d'observation «Somac» développé par Dubé, Pellerin, Déziel et Charrier (1985)[2] utilise aussi une stratégie de codage descriptif dont certains aspects ont été empruntés aux deux systèmes décrits ci-haut. Le

---

[1] La description des méthodes de transcription verbale en synchronie avec les mouvements corporels, faisant aussi partie intégrante de l'instrument de Condon, est omise ici.

[2] Une première version de l'instrument a été opérationnalisée par Déziel (1985); le Somac tel que présenté ici est une version améliorée faisant suite aux recommandations et aux modifications suggérées par cette étude.

construit de la grille a été prévu pour mesurer la variable non verbale du comportement humain lors d'une interaction dyadique en face à face et dans un contexte psychothérapeutique. La codification n'est cependant pas réalisée sur le vif, mais plutôt à partir des enregistrements magnétoscopiques tirés des entrevues dyadiques. Le phénomène observable choisi est le mouvement, plus spécifiquement le mouvement articulaire. La catégorisation des mouvements est d'abord déterminée par les parties du corps (tête, épaules, tronc, bras, avant-bras, mains, doigts, cuisses, jambes, pieds) et en deuxième lieu, par le type d'articulation possible (flexion, extension, abduction, adduction, rotation, circumduction, pronation, supination, protraction, rétraction) dans chaque segment corporel (e.g.: flexion,extension et rotation de la tête): en plus d'être enregistrés par l'inscription d'un code d'identification, ces mouvements articulaires sont repérés dans l'espace en notant les positions des segments corporels sur les axes cartésiens d'un écran vidéo. Quelques autres mouvements particuliers sont aussi codés, mais sans repérage spatial (e.g.: hochements de la tête, balancements de la jambe, pointage de l'index, pianotement des doigts, etc.) à cause de leur rapidité d'exécution. La taxonomie du Somac se caractérise par une micro-décomposition de l'entité «mouvement»; le répertoire obtenu permet une description détaillée, non inférentielle et assez exhaustive des mouvements articulatoires du corps humain; cette transcription microanalytique offre la possibilité d'effectuer ultérieurement aussi bien une analyse structurale qu'une analyse fonctionnelle du mouvement, des gestes ou de la posture des sujets. La méthode d'échantillonnage est continue et complète; un minimum d'intrusion dans le milieu est atteint par des

supports audio-visuels placés discrètement sur le site. L'enregistrement de l'objet d'observation est réalisé par un ou plusieurs observateurs assistés d'un système d'encodage informatisé. Ce système facilite la notation temporelle et spatiale des événements, ou mouvements articulatoires: l'incrustation d'un chronomètre au $^{1}/_{10}$ de seconde sur la bande magnétoscopique permet une lecture de temps assez précise; le repérage manuel des débuts et des fins de mouvements s'opère conjointement avec l'identification des catégories (e.g.: flexion du bras initiée à 01':35":3 et terminée à 01':35":7); le déplacement spatial, i.e. l'amplitude des mouvements, est noté systématiquement par la digitalisation automatique de l'image sur écran aux points d'articulation concernés. Par exemple, après avoir repéré un début de mouvement, l'image est mise en pause et l'observateur enregistre à l'aide de son clavier la partie du corps, le type de mouvement, le moment temporel initial et effectue ensuite la digitalisation sur écran à l'aide d'un curseur pour localiser spatialement le segment concerné; il répète la digitalisation après avoir recherché et enregistré le moment final. A la différence de la grille de Frey et Pool, les unités spatiales sont intégrées au sujet, ou font partie intégrante du sujet (e.g: flexion du bras (Somac) vs bras fléchi à la hauteur de la tête (Berner system)). Les cotes obtenues sont de durées variables et se chevauchent continuellement. On parle donc d'un système multidimensionnel, et à cotes multiples sur certaines dimensions. L'enregistrement spatio-temporel des événements simultanés constitue une méthode de choix pour décrire «objectivement» et «fidèlement» l'objet d'étude; il entraîne cependant plusieurs complications d'analyse puisque les mesures du comportement ont toujours été examinées, par

le passé, après qu'une segmentation, ou un découpage, des événements ait déterminé un nombre d'unités fixes et connues.

Le problème à l'étude dans la présente recherche est ici dévoilé: comment déterminer la valeur métrologique d'un instrument encodant des séquences spatio-temporelles continues et composées d'«événements» simultanés? Certes, plusieurs caractéristiques du système Somac ont pour but de contrôler la qualité des données (e.g.: procédé de codification informatisé, digitalisation directe de l'amplitude du mouvement, chronomètre intégré à l'image, mnémoniques des codes faciles d'utilisation, entraînement systématique des observateurs, etc.) mais, dans une perspective scientifique, la fiabilité réelle du processus observationnel en demeure encore inconnue. Un autre problème se pose: comment concevoir théoriquement la fidélité d'un tel instrument? Les outils statistiques issus des conceptions traditionnelles de fidélité peuvent-ils servir? ou sont-ils adéquats? Une règle acceptée[1] en matière de fidélisation des données stipule que les événements pour lesquels on anticipe un niveau d'accord doivent correspondre à la nature des unités d'observation; ainsi, selon le contexte d'étude[2], les incidences d'accord sont en fonction de plusieurs types de définitions des événements observés. Ce principe, en soi justifiable, implique ici des efforts particuliers menant hors

---

[1] Voir, entre autres, Hartmann (1982), Hollenbeck (1978); Laurencelle (1986); Mitchell (1979).

[2] Par exemple, si le contexte est microanalytique et qu'il implique plusieurs dimensions du même objet d'étude, on aura de nombreuses unités d'observation à considérer individuellement, ou globalement comme composantes de l'accord.

des traditions scientifiques en ce qui concerne la définition des unités de scores. Les items d'un test, ou le codage simple d'un état pouvaient se traiter convenablement par les méthodes psychométriques existantes; mais le codage multidimensionnel avec plusieurs unités à enregistrer, et conçu en plus pour minimiser les biais et les erreurs d'observation, ne peut pas s'ajuster aux théories et méthodes habituelles.

Des propos de Cosnier (1984) expliquent bien l'envergure des problèmes apportés par ce type de recherche et quelques extraits en sont soulignés:

> "Les problèmes méthodologiques posés par l'étude de la «posturo-mimo-gestualité» sont nombreux et complexes: nous nous trouvons en effet à peu près dans une situation analogue à celle où se trouvaient les linguistes à l'époque pré-alphabétique" (p. 8).

Cosnier présente six étapes où ces problèmes se manifestent avec plus ou moins de complexité: le recueil des données (technique), la description des données (extraction de l'information), le traitement des données (définition des objectifs), les corrélations intra et inter-sujets et l'interprétation des résultats. Il poursuit ainsi:

> "La description, le traitement des données sont plus compliqués, ceci pour deux raisons complémentaires qui sont liées aux caractères sémiotiques du canal visuel. (...) La chaîne posturo-mimo-gestuelle est <u>continue</u> dans le temps et <u>tridimensionnelle</u> dans l'espace. La définition des unités sera donc plus délicate: le corps est en état d'émission continue et il

> peut émettre simultanément plusieurs signaux (...). D'autre part la taille et la nature des «unités» choisies peuvent être très variables selon les objectifs poursuivis" (pp.9-10).

Il conclut en spécifiant les contraintes qui conditionnent le choix du chercheur, soit la fidélité du procédé et l'économie du procédé.

La dernière partie, qui suit immédiatement, ébauche une conception nouvelle de la notion de fidélité, répondant ainsi à un besoin de la recherche empirique: celui d'alimenter l'esprit critique et inventif des chercheurs en vue d'acquérir de nouveaux outils scientifiques universels.

Deuxième partie: Proposition d'un modèle de fidélisation des données

Nous avons vu, au chapitre deux, que la façon d'établir la fidélité des données d'observation directe et d'en rapporter les résultats, demeure très controversée. De plus, la "fidélisation" est en général menée par une approche utilisant des données enregistrées selon une structure d'intervalles artificiellement déterminés; cette méthode donne évidemment lieu à une transformation de la "réalité", dont l'ultime but est de s'accommoder aux modèles théoriques déjà existants. Ainsi, des auteurs (Hollenbeck, 1978; Laurencelle, 1986; Sackett et al.,1978) reconnaissent qu'il y a eu négligence à considérer les données enregistrées de façon continue dans la définition des méthodes de calcul de l'accord entre observateurs; même lorsque des méthodes

d'échantillonnage continu sont appliquées, les dimensions comportementales de durée, de séquence, d'espace et d'intensité restent peu explorées, sinon évitées, au niveau des études de fidélité. Sur le plan scientifique, nous nous retrouvons avec des méthodes pratiques qui ne peuvent être appréciées par des modèles théoriques convenables[1].

La transférabilité des concepts de la théorie classique des tests au domaine de l'observation directe est rendue complexe par la différence majeure de leurs objets d'observation; ce n'est qu'au plan opérationnel, soit l'appréciation des résultats d'observation, que les chercheurs appliquent le concept de fidélité. Une approche différente, plus large, du problème de fiabilité du processus observationnel (données, méthodes, juges) apporterait sans doute une conception théorique plus adaptée. Peut-être semblerait-elle imprécise au départ, mais n'aurait-elle pas l'avantage d'alimenter la recherche dans un autre parcours, moins traditionnel que ceux depuis longtemps explorés?... Pourquoi ne pas aborder le problème de l'évaluation des données observationnelles sous un autre angle que celui de comparer les diverses méthodes d'appréciation de la fidélité et de critiquer leur applicabilité? Puisque les efforts de recherche dans le domaine du comportement humain ont encore comme ultime but d'assurer la fiabilité des observations produites et de démontrer leur représentativité d'un "univers" précis, il est indispensable que

---

[1] "L'échec à appliquer les concepts de fidélité à un niveau pratique reflète la confusion théorique à propos du concept de fidélité en recherche observationnelle" (Hollenbeck, 1978 - p. 81).

les méthodes "viables" de repérage des éléments de la "réalité" puissent reposer sur une conception théorique universellement acceptée et adaptée aux nouvelles définitions de l'objet d'observation.

Le système observationnel «Somac», décrit au début de ce chapitre, a été élaboré pour combler les lacunes sus-mentionnées, c'est-à-dire qu'il se distingue sur deux aspects principaux: (1) la rigueur de la méthode d'observation par la description complète, objective et non segmentée des phénomènes; (2) la confrontation aux problèmes d'opérationnalisation de la fidélité des données d'observation directe par l'amorce de méthodes statistiques originales. Ceci ouvre donc la voie au questionnement de la conception théorique de fidélité soutenant la méthodologie de validation d'un tel outil. Nous discuterons ces deux aspects dans les propos qui suivent: le premier portera sur des réflexions ébauchant un point de vue nouveau de la fidélisation d'un processus observationnel; le deuxième présentera des modèles d'opérationnalisation de la «fidélité» des données recueillies par un processus continu et multidimensionnel.

Le phénomène de la «représentativité» dans une conception nouvelle

En somme, deux questions nouvelles se posent: (1) Comment peut-on évaluer un «filtre observationnel» dans son aptitude à représenter la multidimensionnalité d'une portion du «réel»? (2) Quel serait le modèle convenable définissant le processus de filtration de l'observateur dans son repérage des phénomènes? Supposons qu'un bon filtre observationnel possède

des critères bien définis pour facilement percevoir les aspects de localité, temporalité et amplitude de l'objet à représenter, il reste tout de même nécessaire de questionner la pertinence d'observer un objet sous ces angles. Cela revient à dire que tout choix d'observation représente une inférence de l'objet, et que cette inférence peut réduire l'exactitude de la description de l'objet. Comment peut-on estimer ce qu'on enlève, ou ce qu'on néglige, de l'entité «réalité»? Dans la théorie classique des tests, on établit la grandeur de cet écart, «réalité»-«mesure», en estimant la part d'erreur contenue dans les scores des sujets; avec une théorie de l'observation naturelle, on veut aussi estimer le degré de réalité présent dans les notations observées. Toutefois, on ne peut pas démontrer aussi facilement, qu'avec les items d'un test, les cotes des observateurs sont «valables».

Déjà en 1978, Hollenbeck mettait en question l'utilité de rendre ces deux théories isomorphiques: la théorie classique des tests et une «théorie de l'observation naturelle». Il donnait comme argument la différence entre les types d'échantillonnage: les échantillons d'une séquence observationnelle ne sont pas des réplications indépendantes l'une de l'autre, comme le sont en principe les items-réponses à un test; ils sont interreliés. Et cette interrelation des comportements fait partie des finalités mêmes de la recherche observationnelle, puisqu'elle émerge de la structure du comportement observé et qu'elle l'exprime. Hollenbeck concluait ainsi: "La différence entre ces deux théories de mesure a besoin d'être complètement explorée avant que nous puissions totalement comprendre les concepts de

fidélité (p. 82)."

Ainsi, même si certains chercheurs reconnaissent que ces deux contextes d'observation sont tout à fait distincts, personne n'a, à notre connaissance, tenté d'approfondir l'épistémologie de ce modèle contemporain d'acquisition des connaissances, soit l'observation naturelle. Avant de passer à des réflexions en ce sens, nous proposons une description des objets d'observation de ces deux contextes pour mieux faire ressortir ce qui les distingue.

Les évaluations métriques traditionnelles sur le comportement humain s'appliquent à un objet linéaire et unidimensionnel; par exemple, la performance d'un sujet à un test constitue une mesure ou une observation unique avec laquelle le sujet est classé selon une caractéristique précise. Qualifions cette observation d'indirecte, puisque c'est par le biais d'un test qu'on observe le sujet. Nous admettrons aussi que l'observation indirecte implique la fixité des phénomènes étudiés, c'est-à-dire que la «valeur» de la caractéristique évaluée est considérée temporairement stable, au moment où on la mesure. D'autre part, l'objectif actuel de la recherche observationnelle est au niveau du «mieux observer» pour «mieux comprendre»; les méthodes d'observation visent à élargir le «champ d'observation». L'objet d'observation est devenu plus complexe; on tente de repérer des organisations multi-dimensionnelles et séquentielles dans les comportements. Cette observation, dite naturelle (par les éthologistes), suppose, contrairement à l'observation indirecte, la fluidité et le mouvement constant des phénomènes observés. Une autre distinction concerne le rôle de l'observateur: dans l'observation

indirecte, le test est l'instrument d'observation et l'observateur est accessoire; par contre, dans l'autre cas, l'observateur même est juge. Nous avons donc un «instrument» plus complexe devant un objet plus complexe; et chaque réponse (ou observation) du juge est d'une tout autre nature que chacun des scores à un test. Dire de l'observation d'un juge qu'elle est «mesure», nous semble d'ailleurs un abus de l'acception de ce terme. La mesure implique en effet de situer un objet, par rapport à une de ses caractéristiques, sur un continuum; c'est un procédé à la fois descriptif et quantitatif (Bélanger, 1982). La description que le juge fait de son objet nous apparaît comme une manipulation plus ou moins rigoureuse d'un ensemble de caractéristiques provenant à la fois d'un répertoire prescrit et de son expérience personnelle, ceci lui permettant de classer éventuellement ce qu'il observe. A la limite, nous pourrions considérer l'observation naturelle comme l'étape initiale précédant de loin l'observation métrique (ou indirecte); dans ce processus à plusieurs étapes, chaque ensemble d'observations recueillies successivement sur un objet d'étude organise cet objet selon des caractéristiques nouvelles à chaque étape. Ainsi, plusieurs étapes hypothético-déductives amènent graduellement l'objet d'étude à être postulé selon une caractéristique globale et isolée. Alors, le but de notre exposé est d'indiquer l'importance d'orienter la recherche vers des outils statistiques applicables à l'évaluation des données provenant des méthodes d'observation naturelle, ou directe; en d'autres mots, il nous apparaît indispensable de pouvoir apprécier ce processus primaire de l'observation.

Examinons de plus près cette étape primaire de la connaissance des phénomènes en nous situant dans un contexte restreint, soit l'observation du comportement humain. D'abord, nous prenons pour acquis que les éléments «observables» de cette «réalité» sont en mouvement constant et que, tout comme dans l'univers physique de la matière, ils sont de quelque façon en interaction. Supposons que nous voulions étudier le mouvement humain; nous savons que c'est un élément en interrelation avec d'autres éléments d'un «grand ensemble» non défini et nous admettons qu'une partie de nos observations pourra porter sur autre chose que ce que nous prétendons observer, ou que la description de notre objet sera incomplète. Cependant, le phénomène à étudier pourra tout de même être cerné sur les différents aspects observables que nous lui aurons attribués. Nous avons jusqu'ici deux considérations propres à l'observation, soit la «réalité» que l'on veut circonscrire, et l'intention d'observer qui implique un filtrage de l'information potentielle issue de cette «réalité». Nous introduisons maintenant une troisième considération essentielle aux deux premières, soit la fonction de l'observateur. Ce dernier peut être vu comme une «interface» entre le filtre et le flux d'informations extérieures. Par la définition de sa tâche, il utilise le filtre avec lequel il a été entraîné pour repérer certaines dimensions de l'objet à observer. En somme, dans cette tâche, il a à discerner les informations qui peuvent être incluses ou exclues de son filtre. Son action est strictement limitée, au plan théorique, à opérer un tri de ce qu'il perçoit et un classement de ce qu'il peut enregistrer, sauvegarder. Bien sûr, pensera-t-on que l'influence du filtre agit sur la perception qu'a l'observateur

de la réalité; mais rappelons-nous qu'il est lui-même partie de la «réalité» qu'il observe et qu'il est donc également sous son influence. Revenons avec l'exemple du mouvement humain comme objet d'observation pour illustrer ces influences; lorsque l'observateur décode une suite de mouvements corporels tels que définis par son filtre, il utilise sûrement sa propre connaissance de se mouvoir, dans le repérage des mouvements chez un autre sujet. Cela démontre comment tout est interrelié et comment il devient illusoire d'exiger la précision absolue de tels «instruments», l'observateur et le filtre. Nous n'élaborons pas plus longuement sur ces questions puisqu'elles ont été souvent commentées avec les diverses notions de biais dans l'observation.

Considérons à nouveau le rôle de l'observateur, lequel selon nous n'est autre qu'un élément d'un nouveau système «réalité-filtre-observateur», que nous appellerons «processus observationnel», cette appellation se distinguant toutefois de celle de Laurencelle (1986) parce que celle-là excluait la fonction de l'observateur; donc, dans sa fonction de jonction entre deux supra-systèmes (réalité et filtre), l'observateur doit maintenir un mouvement constant entre le flux d'informations (réalité) et le filtrage des informations (instrument d'observation ou filtre) pour être opérationnel. S'il y a échec à cet échange constant de l'information, des dysfonctionnements de l'observation apparaîtront; nous présentons deux exemples extrêmes de désorganisation: (1) l'observateur utilise le filtre de façon «obsessive» et nulle information de l'extérieur ne peut être classée; (2) l'observateur laisse trop pénétrer d'informations de la «réalité» et il devient sursaturé; devant

trop de choix difficiles, il classe au hasard. Enfin, pour mieux expliquer la désorganisation de notre interface dans ce point de vue systémique, posons quelques hypothèses sur le fonctionnement de l'observateur dans son interaction avec d'autres systèmes ou sous-systèmes: par exemple, dans un système «recherche observationnelle», le sous-système «chercheur», qui emploie un observateur et en assume éventuellement l'efficacité, pourrait négliger de vérifier le travail d'observation; ou bien le sous-système «observateur» pourrait manquer de rigueur dans son travail, ou même éviter d'en faire évaluer certaines parties. Nous avons ici des exemples d'un fonctionnement relâché, mais, à l'inverse, un fonctionnement rigide et contrôlant pourrait aussi avoir un impact négatif sur la qualité du travail de l'observateur. Par ces hypothèses de dysfonctionnement, nous voulons expliquer une forme de «pathologie» d'observation apparaissant comme suit: avec le temps, le jugement de l'observateur devient moins rigoureux et l'observateur se «décalibre»; il produit alors des biais idiosyncratiques systématiques, caractérisant la détérioration de l'«interface».

La fonction du juge, dans son appellation «interface», peut aussi être expliquée avec les trois propriétés qui lui sont attribuées en informatique. Ainsi, en sciences de l'information, une interface fonctionnelle possède les qualités suivantes: linéarité, étendue et calibration. Dans notre théorie de l'observation, le juge repère les transformations dans son objet d'observation et les «conserve» (enregistre) en données invariantes; dans cette fonction, il opère des correspondances consistantes entre le champ d'observation et le

système d'observation (filtre), ayant par le fait même exclu certaines variables de ces transformations. Dans ce contexte, l'analogie de l'«étendue» s'exprime dans la «sensibilité» du juge, ou sa «tolérance» à repérer les signaux (phénomènes) dans le champ d'observation; la «linéarité» trouve définition dans la reconversion des signaux, ou leur reproduction dans le même spectre de fréquences et d'intensités, en particulier dans l'attribution consistante des mêmes codes aux objets de mêmes types. Nous constatons de ce fait qu'un rapport optimal doit être maintenu entre «étendue» et «linéarité», de façon à ce que la proportion tende vers l'unité. De plus, le juge opère ses classements suivant un calibre; théoriquement, sa calibration se maintient après chaque mesure d'un événement, i.e. après chaque cycle comportant l'évaluation d'un signal et, s'il y a lieu, l'enregistrement d'un classement. Avec ce modèle, nous pouvons aussi redéfinir les dysfonctionnements d'observation comme suit: (1) si le juge cesse d'être linéaire, les phénomènes perçus perdent de leurs proportions réelles; cette déformation des signaux mène à deux tendances extrêmes: la perte d'étendue, empêchant même les classements, et le gain d'étendue, entraînant trop de classements, dont plusieurs au hasard; (2) le désajustement de la calibration du juge survient lorsqu'à chaque cycle «évaluation/classement», le calibre se trouve modifié; il n'y a pas un retour à l'état originel de calibration.

Nous abordons maintenant une description plus concrète de notre perspective systémique appliquée au processus observationnel[1]. Nous

[1] La figure 1 plus loin fournit une représentation schématique de cet ensemble.

supposons que l'objet étudié se schématise comme une ligne brisée montrant tantôt des plateaux, tantôt des pignons. Nous appelons cela la «réalité». Nous plaçons l'observateur que nous avons auparavant décrit comme une «interface», avec son filtre lui permettant de percevoir une portion de ces manifestations (voir figure 1, comme schéma illustratif). Sa décision d'étiqueter une manifestation par une catégorie de son répertoire (filtre) peut être interprétée ainsi: la manifestation dans l'objet d'étude suscite à un certain moment l'intérêt perceptif chez le juge; la réaction résulte éventuellement dans l'organisation d'une nouvelle perception qui sera classée dans un élément du filtre. La «chaîne d'émission de la réalité» étant en mouvement constant, une manifestation peut tantôt baisser d'intensité jusqu'à disparaître, demeurer au même niveau, ou augmenter de telle façon que le juge ait à décoder ce qui se passe; donc, on peut supposer que le phénomène montre un mouvement (hausse d'intensité, contraste accentué d'une manifestation d'un type donné par rapport au type antécédent) tel qu'un «seuil de tolérance» soit atteint chez le juge et que ce dernier perçoive un changement dans le «continuum» observé.

Enfin, la décision d'un juge de réagir à une manifestation peut être interprétée comme l'atteinte de son niveau maximal de tolérance relativement à une définition spécifique de l'objet observé; et même s'il s'exerce une interaction constante entre le juge et la chaîne de manifestations, cette dernière n'offre suffisamment de stimulation qu'à certains moments spécifiques.

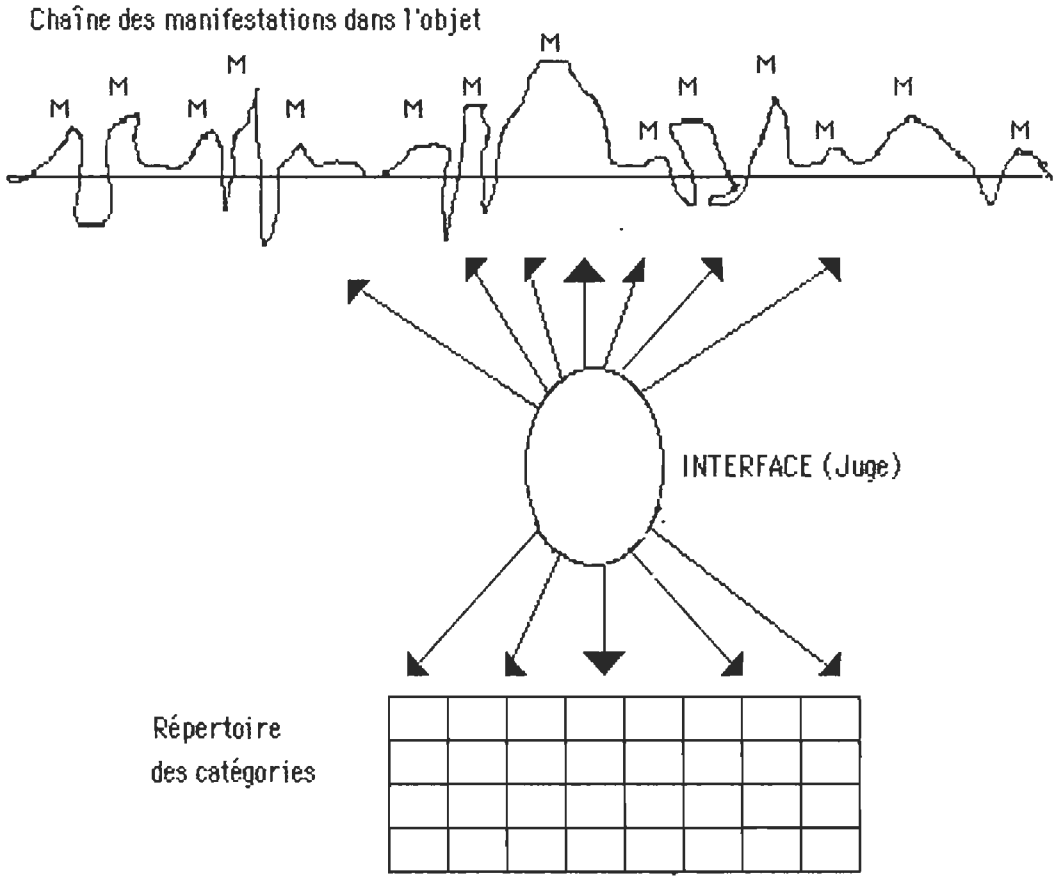


Fig. 1 - Représentation des composantes du processus observationnel.

Articulons les particularités de ce modèle dans un langage symbolique, où les termes seront définis comme suit:

'**M**' = manifestation (fraction observable du réel, en dimension 'r')

'$I_J(M)$' = intensité de la manifestation **M** dans la catégorie **j**

'S_J' = seuil absolu d'amission de catégorie **j**

'$S_{JJ'}$' = seuil différentiel des catégories **j** et **j'**

D'autres aspects relatifs à la systématisation de l'observation vont conditionner les éléments ci-haut définis. Par exemple, revenons aux procédures de codage impliquant l'emploi de plusieurs catégories pour décrire une même manifestation (codage multiple), et l'observation de plusieurs manifestations, ou dimensions différentes du même objet (codage multidimensionnel), et nous devons alors parler de la «spécificité» des seuils. Cette spécificité peut être considérée sur deux plans: les seuils se situent aux points de transition entre l'absence et la présence d'un phénomène (seuil absolu), et entre l'absence et l'existence d'une différence entre deux types de manifestation (seuil différentiel). De plus, les seuils différentiels ont la propriété de s'adapter aux phénomènes, c'est-à-dire qu'ils s'ajustent aux intensités des phénomènes ainsi qu'aux divers contextes d'occurrence.

Toutefois, à part la complexité de la tâche d'observation lorsque plusieurs catégories, ou plusieurs manifestations sont recherchées, nous pouvons supposer que le processus d'étiquetage d'une manifestation comporte invariablement la nécessité d'une coïncidence entre un seuil et l'intensité de la manifestation. Ainsi, mieux les catégories de classement seront définies clairement, plus les seuils établis après l'intégration des catégories (ou du filtre) seront optimaux et discriminatifs, prévenant ainsi des choix au hasard à cause des perceptions indifférenciées.

A présent représentons-nous ce modèle de l'observation dans quelques situations de codage:

En codage multiple:

Un classement spécifique s'effectue lorsque le seuil d'un juge est réduit, ou qu'un seuil absolu est atteint, par la perception d'une intensité optimale dans une manifestation et lorsqu'il y a pairage entre le seuil ainsi atteint et une catégorie pour cette intensité. Si l'intensité est insuffisante en rapport au seuil, le juge ne perçoit pas de changements. Dans cette forme de codage, plusieurs catégories peuvent définir une même manifestation et la tâche du juge consiste à classer toutes les intensités perçues rencontrant chacune un seuil propre.

La règle, en résumé, peut s'exprimer comme suit:

Chaque **j** pour lequel $I_J(M) > S_J$ est admis. (A_1)

En codage simple (ou codage multidimensionnel simple):

L'identification de la meilleure description pour catégoriser chaque manifestation suppose des stratégies de comparaison et d'élimination; il y a recherche de l'intensité dominante pour réduire la tolérance d'un juge, ou atteindre un seuil différentiel. Avec cette forme de codage, le juge doit aboutir à une seule valeur (ou catégorie) pour chaque manifestation observée. Nous présumons qu'une stratégie prévaut lors de ces opérations, c'est-à-dire que le juge cherche à trouver la catégorie dominant les autres.

En résumé, une opération de classement consiste à:

Trouver le j pour lequel $C_{JJ'} = I_J(M) - I_{J'}(M) > S_{JJ'}$ (B_1)

et $I_{J'}(M) - I_{J''}(M) \not> S_{J'J''}$ (B_2)

pour tous $j' \neq j$ et $j'' \neq j'$

Supposons que cette première stratégie fasse échec, nous pouvons conceptualiser l'échec de deux façons: (1) le juge n'arrive à classer aucune valeur j parce qu'il n'y a pas d'intensité dominante et qu'un seuil différentiel particulier n'est pas dépassé; (2) le juge aboutit à plusieurs valeurs j parce que plusieurs seuils différentiels sont atteints. Une autre approche doit alors être employée, de façon à:

Trouver la catégorie majoritaire,

telle que $C_{JJ'} > C_{J''J'}$, pour tous les j'' (B_3)

Dans sa fonction d'«interface», le juge utilise idéalement la stratégie donnant priorité à une catégorie spécifique, c'est-à-dire, comme le montrent les formules B_1 et B_2, celle pour laquelle un seul seuil différentiel a été dépassé. Lorsque le classement est plus difficile, une autre stratégie consiste à comparer toutes les paires de catégories pour obtenir une catégorie majoritaire (B_3), c'est-à-dire celle dont le seuil différentiel a été rejoint de plus près (premier type d'échec), ou celle dont le seuil différentiel a été le plus fortement atteint (deuxième type d'échec). En d'autres mots, par cette deuxième stratégie, le juge réussit à choisir la catégorie pour laquelle sa

tolérance est le plus fortement réduite. Si cette deuxième stratégie échoue à fournir l'évidence d'une seule catégorie, le juge effectuera une modification de son rapport «étendue/linéarité», opération risquant d'introduire des pathologies d'observation: (1) Lorsque l'échec consiste à ne trouver aucune intensité dominante, le juge cherche à gagner de l'«étendue»; il pourrait alors baisser ses seuils à un point tel que de fausses catégories soient incluses. (2) Si la difficulté est d'avoir trop de catégories dominantes, le juge tente de perdre de l'«étendue»; il élève ses seuils de façon à éliminer l'influence de certaines catégories et il y a danger d'exclure même de bonnes catégories.

Par conséquent, une condition d'opérationnalisation idéale des seuils de l'«interface» exige, sauf toute exception, que pour un contenu de référence donné, un juge dispose de valeurs optimales de seuils absolus, ou de seuils différentiels; les seuils du juge sont S_J ou $S_{JJ'}$ et les seuils optimaux prescrits par le répertoire de référence sont σ_J ou $\sigma_{JJ'}$. La série (σ_1, σ_2,..... σ_R) et (σ_{12}, σ_{13},..... $\sigma_{R-1,R}$) constitue alors le calibre du juge.

Cependant, dans le cas du codage simple, plus le nombre de manifestations à classer augmente, plus la série de seuils différentiels s'allonge et la tâche de classement se complexifie; c'est que le calibre du juge comporte moins de démarcations entres ses composantes, accroissant ainsi la difficulté à atteindre un seuil électif. Le codage multiple est donc à préconiser puisque le danger de faux classements y

est faible, ramenant ainsi les possibilités d'erreurs aux seules omissions.

Enfin, l'adéquation d'un seuil optimal par rapport à une catégorie particulière devrait correspondre à la qualité de la reproduction obtenue par l'«interface» pour une intensité donnée. La bonne correspondance entre un seuil optimal et le seuil appliqué à une catégorie équivaut, en fait, au rapport optimal «étendue»/«linéarité», lequel constitue la qualité de «calibration» de l'«interface». La constance de ce rapport fait état de l'aptitude de l'interface à conserver son calibre après chaque classement.

Nous voici à l'étape critique de définir un moyen d'apprécier la qualité d'application de ce modèle. Nous savons déjà par l'observation métrique que toute évaluation des fonctions humaines, aussi objectives ou manifestes qu'elles peuvent être considérées, implique une inférence; par une évaluation métrique, nous sélectionnons un objet spécifique observable comme l'indicateur d'un concept général non observable. L'exigence d'outils ou de mesures démontrés fidèles, stables et consistants peut mieux être comprise dans ce contexte. Cependant, c'est effectivement à cause des conséquences appauvrissantes de ce procédé inférentiel que les méthodes d'observation naturelle ont été développées. Il nous apparaît par conséquent inapproprié de recourir aux mêmes méthodes de vérification puisque nous avons expliqué qu'il est question d'un tout autre processus d'observation.

Comme Laurencelle (1986) l'avait déjà énoncé, les préoccupations d'évaluation dans un contexte d'observation naturelle comportent deux seuls aspects: celui de la véracité des données et celui de la performance du juge. Dans un premier temps, il s'agit de concevoir une évaluation de notre «interface», puisque c'est celle-ci qui effectue le décodage des événements. Qu'est-ce qui peut donc nous permettre de nous assurer de l'adéquation de l'«interface», i.e. de la bonne performance du juge? Soulignons d'abord les réponses implicites apportées par le modèle que nous venons d'élaborer. Par exemple, on peut déduire que les procédures d'entraînement d'un observateur doivent être bien systématisées de façon à ce qu'un juge puisse se définir des seuils précis et appropriés. L'utilisation d'un répertoire de catégories adéquat, i.e. avec des définitions claires, objectives et pertinentes pour chaque manifestation que l'on veut circonscrire, constitue pour le juge une source de référence facilitant l'établissement de ses critères de classement, ou seuils.

La seconde question est: comment apprécier la «véracité» de nos reproductions obtenues par l'observation naturelle? Le principe même de l'observation naturelle, impliquant l'observation simultanée de plusieurs caractéristiques de l'objet, constitue, à notre avis, un ensemble de conditions qui lorsque remplies ajoutent à la fidélité d'une reproduction. En d'autres mots, si deux juges s'accordent sur la reproduction de plusieurs dimensions d'un même phénomène, enregistrées simultanément, cela augmente la probabilité que le phénomène soit «réel». Cette hypothèse répond d'ailleurs à

une perpective systémique de l'observation: plus il y a d'éléments d'un ensemble qui sont observés, plus on augmente les chances de bien représenter le phénomène. La conséquence est qu'on augmente aussi la difficulté de reproduire, réduisant ainsi les chances d'accord parfait.

En fait, nous venons d'expliquer que la systématisation de l'observation doit optimaliser la possibilité que les seuils S_J ou $S_{JJ'}$ correspondent aux seuils optimaux σ_J ou $\sigma_{JJ'}$ pour chaque catégorie de contenu du répertoire. Aussi, nous supposons que l'observation multidimensionnelle comporte en soi un premier aspect de la représentativité des phénomènes parce que plusieurs dimensions du «réel» sont considérées pour un même objet ou phénomène. Toutefois, ces simples constatations des sous-entendus de l'observation systématique ne donnent pas une consistance suffisante à nos méthodes d'évaluation d'un tel processus. Donc, comment évaluer l'adéquation d'un seuil à une catégorie pour être en mesure d'assumer que les intensités perçues sont véritablement le produit de la manifestation recherchée? Nous voilà en train de supposer une nouvelle conception de la «fidélité» du processus observationnel s'énonçant comme suit: "lorsqu'un juge applique adéquatement[1] les seuils correspondant parfaitement aux catégories de son répertoire, il est fort probable qu'il attribue les bonnes valeurs, ou catégories, c'est-à-dire qu'il effectue un bon classement des manifestations qui ont interpelé ses seuils. Il en résulte que cette catégorisation, ou

---

[1] Cette adéquation se définit par l'application la plus rapprochée des seuils optimaux σ_J et $\sigma_{JJ'}$, c'est-à-dire que sa sensibilité aux intensités des manifestations est optimale pour le classement de ces dernières.

classement, mènera à une reproduction comparable des événements observés".

Dans cette optique, on peut théoriquement supposer que la "proportion de non-représentativité" correspond à l'écart qui s'est inséré entre un seuil appliqué et son optimum initial ($S_J - \sigma_J$, $S_{JJ'} - \sigma_{JJ'}$). Cet écart se manifeste donc, aux extrêmes, dans la forme des pathologies d'observation présentées plus tôt. On aura trois types d'écarts que nous décrivons comme suit:

1) Une élévation des seuils amène un écart de type $S_J > \sigma_J$; cet écart empêche l'intensité d'une manifestation de susciter la réduction de la tolérance du juge; il y a donc une difficulté accrue à classer dans une catégorie. En définitive, on observe une baisse de la capacité de classer. Dans ce cas, des taux acceptables de fidélité intra-juge et inter-juges peuvent quand même être obtenus pour les classements effectués.

2) Un abaissement des seuils entraîne un écart de type $S_J < \sigma_J$ (ou $S_{JJ'} < \sigma_{JJ'}$) et résulte dans un état constant d'augmentation de la tolérance; ceci donne un pouvoir exagéré à l'intensité d'une manifestation et donc, augmente la facilité à classer dans une catégorie. Le premier mouvement d'abaissement des seuils faisant suite à la persistance d'une ambiguïté au niveau de certaines catégories à coder, va permettre au juge de choisir les bonnes catégories et de maintenir sa compétence. Cependant, si la tendance à diminuer les seuils constitue une décalibration accentuée du juge, les classes deviennent surdéterminées et le classement dans certaines catégories s'effectue au hasard; il y alors une baisse de compétence du juge laquelle se

reflètera par la diminution de l'accord inter-juges, ou par.une certaine part d'accords attribuables au hasard.

3) Des seuils fluctuants font varier l'écart ($\mathbf{Var(S_J) > 0}$ ou $\mathbf{Var(S_{JJ'}) > 0}$), de telle sorte que c'est l'inconsistance dans le choix des catégories; la méthode d'attribution des catégories est flottante. C'est la décalibration de l'«interface». Des évaluations de ce type de classements afficheraient une faible fidélité intra-juge.

Il est possible, bien entendu, qu'un seul, ou que des combinaisons de ces trois types d'écarts fassent l'objet du classement inadéquat, ou non-représentatif pour une même séance d'observation d'un juge.

Connaissant ce que comportent «représentativité» et «non-représentativité» avec notre modèle nouvellement ébauché, nous constatons que des méthodes d'appréciation déjà utilisées, ou déjà suggérées par certains auteurs, conviennent à cette conception du processus observationnel. Entre autres, considérons certains types d'approches de fidélité, présentés au chapitre deux, comme des évaluations appropriées, du moins dans ce qu'elles apportent en information sur l'utilisation d'un répertoire de catégories (filtre) par un juge:

1) La fidélité par catégorie permet d'établir le profil de «représentativité» des classements, entre deux juges, ou chez un seul juge, pour chaque catégorie du répertoire, donc comment chaque seuil est opérationnel. On peut ainsi dégager quelles catégories présentent de l'ambiguïté, ou quelles sont les inadéquations des seuils, et pour quel juge en

particulier (ou à quel moment spécifique pour un même juge). Cette approche s'applique aussi bien dans les deux types de codage, simple ou multiple.

2) La fidélité par codeur, ou par juge, permet d'établir la performance moyenne d'une «interface» par rapport à l'ensemble de ses classements, ou à certains en particulier.

3) La fidélité par événements continus multiples permet d'établir les niveaux de reproductibilité des différentes dimensions classées simultanément (i.e.: quelles sont les dimensions qui influencent négativement ou positivement le niveau d'accord inter-juges ou intra-juge?). Rappelons qu'une seule, ou plusieurs manifestations peuvent faire l'objet d'une dimension et que leur classement peut s'effectuer sous forme de codage simple, ou multiple. Cependant, nous pouvons constater à ce point-ci la complexité d'évaluer la représentativité de tels classements puisque l'objet d'observation est conservé dans sa globalité, contrairement à l'observation unidimensionnelle où l'objet représente une dimension séparée.

4) La fidélité par consensus permet d'établir le degré de correspondance entre les classements de différents juges et d'inférer s'il y a concordance entre les seuils des juges, ou égalité des calibres.

5) La fidélité par critère de vérité nous donne le degré d'intégration des catégories (ou d'adéquation des seuils) par un juge. C'est donc aussi un moyen de vérifier l'état de la «calibration» absolue du juge.

Ces approches de fidélité constituent seulement une mention partielle des méthodes applicables dans le processus d'évaluation des données obtenues

dans le contexte de l'observation naturelle. D'autres méthodes statistiques sont aussi recommandables, tel le calcul du taux d'occurrence d'une manifestation et de la probabilité que certaines catégories soient choisies, et elles viennent compléter l'investigation de la valeur des données. Nous considérons cependant que ces discussions sont l'objet d'une étude opérationnelle, laquelle s'écarte de l'objectif d'élaborer une conceptualisation comme nous l'avons annoncé dans cette section. Aussi, la section qui suit traitera de l'opérationnalisation d'un modèle d'évaluation de la représentativité des données résultant de l'observation multidimensionnelle, ceci se présentant comme la concrétisation des concepts nouveaux présentés ci-haut.

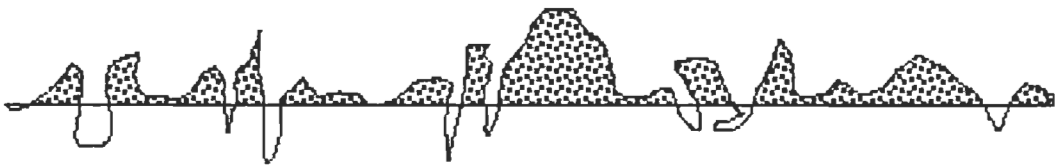
Troisième partie: Différentes propositions d'opérationnalisation de la fidélité

La section précédente nous a introduit dans de nouvelles conceptualisations de la tâche d'observation et de la catégorisation de l'objet d'observation. Ces aspects ont une influence sur notre conception de la reconstruction observationnelle de la «réalité». Cette reconstruction est par ailleurs conditionnée par le modèle perceptuel avec lequel on veut reproduire les phénomènes. Assurément, lorsqu'on opte pour une reproduction à la fois continue et multidimensionnelle, on complexifie, et la tâche d'observation, et l'objet d'observation. Toutefois on augmente aussi la probabilité que les

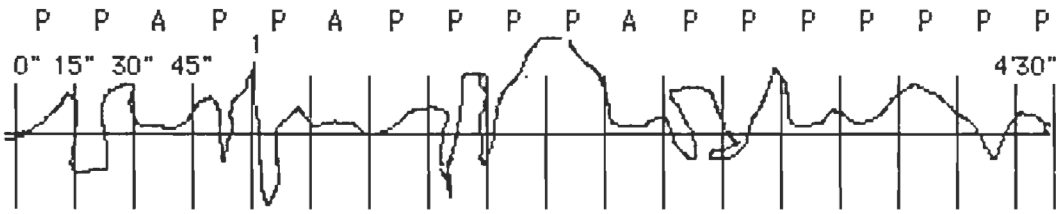
reproductions soient «vraiment» représentatives des phénomènes observés. Pour illustrer cette prémisse, retournons au graphique de la section précédente où nous concevions les manifestations comportementales dans une chaîne accidentée montrant un processus d'émission en activité constante. Supposons que les pignons sur cette chaîne soient des événements, ou manifestations, avec diverses significations et intensités. Admettons de plus qu'une reproduction complète de ces phénomènes impliquerait qu'une multitude de points (ou de dimensions des phénomènes) soient observés à l'intérieur de ce tracé, de façon à obtenir une superposition du tracé des observations sur le tracé «réel». En comparant l'observation multidimensionnelle avec les autres stratégies d'observation comme le scrutage par segments (ou intervalles), et la recherche sporadique d'événements épars, nous sommes amenée à constater, avec ces dessins, que la reproduction des classements unidimensionnels ne serait que très peu représentative de la chaîne des dimensions «réelles». Ces reproductions seraient plus ou moins «partielles», ou «approximatives», selon la stratégie choisie. Les dessins de la page suivante mettent en évidence ces distinctions entre les divers protocoles de reproduction des phénomènes.

Notre discussion a pour but de souligner la complexité d'évaluer le degré de coïncidence entre les protocoles de reproduction de plusieurs juges lorsque cette reproduction implique une catégorisation multidimensionnelle et continue. Même si la reconstruction concordante d'un même objet dans ses nombreuses dimensions observées assure une reproduction plus fidèle de l'objet original, il est autant nécessaire d'apprécier la vraisemblance de cette

DESSIN 1: Reproduction multidimensionnelle continue.



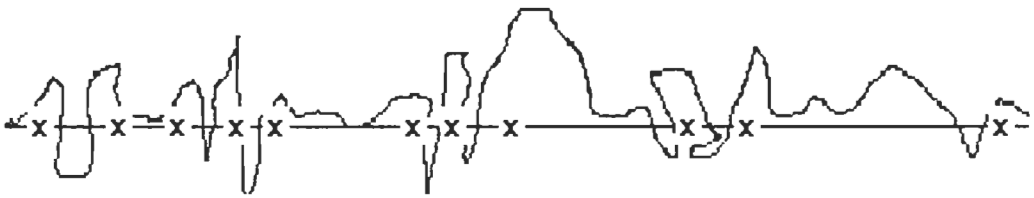
DESSIN 2: Reproduction unidimensionnelle par intervalles.



P=présence de la manifestation dans chaque intervalle de 15"

A=absence

DESSIN 3: Reproduction unidimensionnelle par faits pêle-mêle.



x=occurrence d'une manifestation

Fig. 2 - Illustration de trois types de reproduction selon les méthodes de codage, après qu'un juge ait classé les manifestations perçues.

reproduction. Sachant que les diverses dimensions de l'objet ne sont habituellement ni fixes ni parallèles, le premier problème réside dans la définition même de cette coïncidence, ou de cette superposition. De quoi sera constituée notre unité d'accord? Et quels critères justifieront cet accord?

A ce stade de notre exposé, il est plus utile de se baser sur un modèle concret pour offrir des réponses aux questionnements ci-haut introduits. Nous avons donc choisi un instrument d'observation présentant les caractéristiques de multidimensionnalité et de continuité du codage: il s'agit du système Somac que nous avons décrit auparavant. Ainsi la catégorisation des mouvements du corps humain recoupe les dimensions suivantes: la partie du corps spécifiquement en mouvement, le type d'articulation de la partie du corps désignée, la localisation temporelle et spatiale des mouvements articulatoires, et l'intensité ou amplitude des mouvements. Les classements effectués par ce système sont donc de durées variables, se chevauchant sur différentes dimensions non-sectorisées dans le processus de catégorisation. Alors le problème à affronter réside dans la continuité et la simultanéité des unités de classement. Comment comparer des classements à composition variable?

Nous précisons que notre but ici exclut une présentation complète du processus de fidélisation des données d'observation multidimensionnelle. Ces procédures d'évaluation exigent des applications informatiques de grande envergure; ces dernières constitueraient à elles seules l'objet d'une nouvelle recherche.

Ainsi, nous élaborerons plutôt diverses facettes d'opérationnalisation de la fidélité et elles seront abordées séparément. En résumé, nous considérerons d'abord une fidélité portant sur la quantité de mouvements repérables chez un sujet. Dans cette première section, nous y définirons les diverses constituantes de l'accord, tel que prévu par le système Somac. Une deuxième

section portera sur un autre aspect de fidélité concernant la conception d'une matrice globale pour les classements d'un juge lors d'une séance de codification donnée. La troisième section traitera d'une nouvelle forme de validité, considérant que l'objet d'observation du Somac se prête à une procédure de reconstruction des déplacements observés. Enfin, la dernière section fournira les résultats d'une première étude de fidélité, réalisée cependant de façon très partielle, laquelle apporte un aperçu de la fidélité globale des données du Somac obtenues lors d'une première expérimentation de l'instrument.

Fidélité sur la quantité de mouvement observée

Le Somac comporte au total 220 mouvements possibles par la combinaison des 77 types de mouvements différents à identifier en fonction de la latéralité et des onze parties du corps. Les mouvements sont cependant catégorisés selon trois distinctions comme suit: les mouvements de catégorie nominale, les mouvements de catégorie nominale avec leur orientation et les mouvements «normaux» avec enregistrement de leur déplacement spatial (voir Tableau 6). Pour définir une incidence d'accord, nous devons au préalable identifier toutes les dimensions impliquées dans l'objet d'observation. Notre unité de classement est multidimensionnelle et inclut les aspects suivants: la partie[1] du corps en mouvement; la latéralité; le type[2] de mouvements; le temps

---

[1] Il est plus fréquent d'observer plusieurs parties du corps simultanément en mouvement.

[2] Dans plusieurs cas, il y a plus d'un type de mouvements à identifier simultanément pour une même partie du corps. Nous rencontrons ici le principe du codage multiple.

du début du mouvement; le temps de la fin du mouvement; et pour les mouvements «normaux», deux points de digitalisation aux extrémités du segment corporel en mouvement pour le temps initial; deux autres points de digitalisation à la fin du mouvement, c'est-à-dire au temps final. La figure * 3 illustre l'ensemble des points de digitalisation dans les segments corporels. Nous représentons ci-après des modèles de codification pour les trois types de catégorisation des mouvements:

(1) **TE A HOFE 1 28 32 5 1 28 33 1**

Partie du corps = Tête
Latéralité = Aucune
Mouvement = Hochement en flexion/extension
Début = 1 hre 28 min. 32 sec. $^{5}/_{10}$
Fin = 1 hre 28 min. 33 sec. $^{1}/_{10}$

(2) **EP D PTRE 0 51 25 8 0 51 26 0**

Partie du corps = Epaule
Latéralité = Droite
Mouvement = Protraction à rétraction
Début = 51 min. 25 sec. $^{8}/_{10}$
Fin = 51 min. 26 sec.

(3) **BR G ABDU 2 10 17 3 16080 10872**
2 10 17 9 16279 16563

Partie du corps = Bras
Latéralité = Gauche
Mouvement = Abduction
Début = 2 hres 10 min. 17 sec. $^{3}/_{10}$
Fin = 2 hres 10 min. 17 sec. $^{9}/_{10}$
Digitalisation = Coordonnées **x,y** aux points d'articulation épaule et coude droits:
160,80 108,72 (début du mouvement)
162,79 165,63 (fin du mouvement)

Ces exemples illustrent comment les mnémoniques utilisées dans la codification sont simples ainsi que directement associées à chaque spécification du codage. De plus, l'enregistrement des codes du Somac s'effectue sur micro-ordinateur et le programme prévoit des vérifications de codes erronés. Donc tout enregistrement incorrect au niveau des mnémoniques, ou au niveau du temps, et ne répondant pas aux spécifications du programme, sera signalé et devra être ré-inscrit.

Nous disposons maintenant des éléments pour définir ce qui constitue la «fidélité sur la quantité de mouvement observée»; il s'agit de trouver pour chaque paire de codeurs le nombre total de mouvements identiques qu'ils ont repérés dans les mêmes intervalles de temps et, dans le cas des mouvements «normaux», qu'ils ont localisés aux mêmes coordonnées cartésiennes, par rapport à tous les mouvements classés par l'ensemble des codeurs. Les classements des codeurs sont d'abord comparés par rapport à leur précision d'identification et à leur précision temporelle, et s'il y a lieu, par rapport à leur précision spatiale, c'est-à-dire aux positions segmentaires localisées en coordonnées cartésiennes. Initialement, deux critères sont nécessaires pour justifier la coïncidence d'un classement: la catégorisation du mouvement (incluant partie du corps, latéralité, type de mouvement et s'il y a lieu, direction du mouvement) et l'intervalle temporel où il a été repéré.

Tableau 6

Parties du corps, dimensions articulaires, catégories de mouvements et échelle

N.B.: Les lettres majuscules en caractères gras représentent les symboles mnémoniques utilisés dans la codification.

Partie du corps	# de dimensions codées	Dimension	Type de mouvement			Echelle	
			Nominal	Nominal + direction	Normal & digitalisation	Temps	Amplitude
TEte	4	**HO**chements en: **F**lexion/Extension, **F**lexion **L**atérale, **RO**tation	X			X	
		FLeXIon/**EXTE**nsion			X	X	X
		FLexion l**AT**érale			X	X	X
		ROTAtion			X	X	X
EPaules	2	**EL**évation/**NE**utre/ **AB**aissement		X		X	X*
		ProTraction/ **NE**utre/**RE**traction		X		X	X*
TRonc	4	**F**lexion/Extension **DOR**sale	X			X	
		ROtation **D**roite/ **N**eutre/**RO**tation **G**auche		X		X	X*
		FLeXIon/**EXTE**nsion			X	X	X
		FLexion l**AT**érale			X	X	X
BRas	2	**FLeXI**on/**EXTE**nsion			X	X	X
		ADDUction/ **ABDU**ction			X	X	X
AVant-bras	4	**Pro**Nation/**NE**utre/ **SU**pination		X		X	X*
		Rotation en **AD**duction Interne/Externe			X	X	X
		Rotation en **AB**duction Interne/Externe			X	X	X
		FLeXIon/**EXTE**nsion			X	X	X
MAins	4	**CIRC**umduction	X			X	
		SECOuement	X			X	
		FLexion/**NE**utre/ **EX**tension		X		X	X*
		ADduction/**NE**utre/ **AB**duction		X		X	X*

(suite - Tableau 6)

Parties du corps, dimensions articulaires, catégories de mouvements et échelle

N.B.: Les lettres majuscules en caractères gras représentent les symboles mnémoniques utilisés dans la codification.

Partie du corps	# de dimensions codées	Dimension	Type de mouvement: Nominal	Type de mouvement: Nominal + direction	Type de mouvement: Normal & digitalisation	Echelle: Temps	Echelle: Amplitude
DOigts	7	**CR**oiser les doigts/ **NE**utre	X			X	
		COntact des doigts/ **NE**utre	X		X		
		POinter de l'index/ **NE**utre	X			X	
		PIANoter	X			X	
		Contact partie du Corps/**NE**utre	X			X	
		FLexion/**NE**utre/ **EX**tension		X		X	X*
		ADduction/**NE**utre/ **AB**duction		X		X	X*
POuces	3	**OP**position/ **NE**utre	X			X	
		CIRCumduction	X			X	
		FLexion/**NE**utre/ **EX**tension		X		X	X*
CUisses	2	**AD**duction/**NE**utre/ **AB**duction		X		X	X*
		FLEXIon/**EXTE**sion			X	X	X
JAmbes	4	**BAL**Ancement	X			X	
		ROtation Interne/ **Ne**utre/**RO**tation Externe[1]		X		X	X*
		FLEXIon/**EXTE**nsion			X	X	X
		ROTAtion			X	X	X
Pieds	3	**CIRC**umduction	X			X	
		FLexion/**NE**utre/ **EX**tension		X		X	X*
		EVersion/**NE**utre/ **IN**version		X		X	X*

[1] Ce mouvement s'effectue au niveau de la cuisse, mais il est repéré au niveau de la jambe.
* Mouvements où l'amplitude est enregistrée de façon nominale mais non de façon quantitative.

Points de digitalisation indiquant le déplacement pour (6) segments corporels

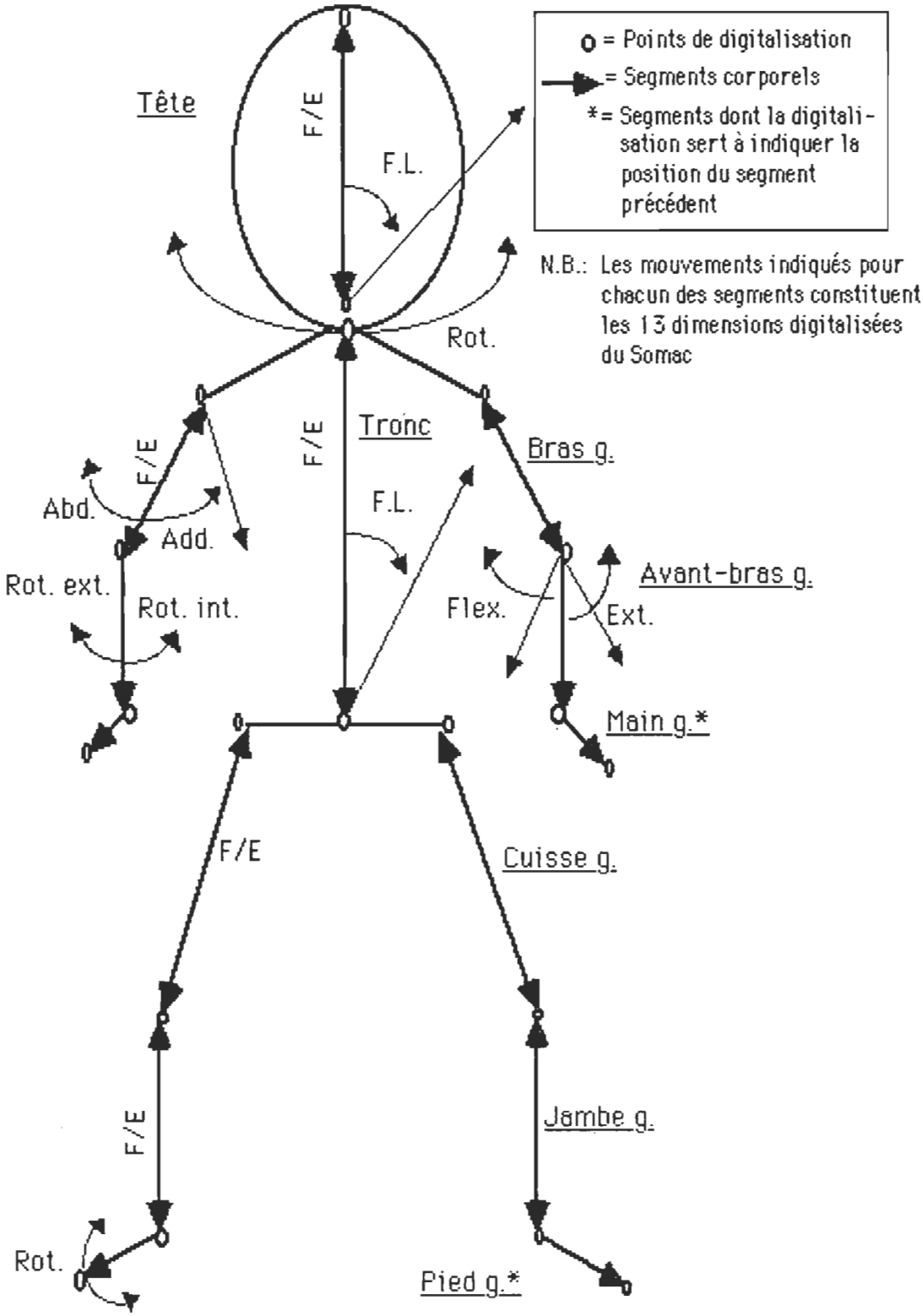


Fig. 3 - Illustration des segments digitalisés et des mouvements du Somac.

Une fois que les catégorisations ont été bien assorties, la procédure d'appariement des classements suppose un recouvrement parfait, de type '1' ou '0', des intervalles temporels pour obtenir une coïncidence entre deux classements de différents codeurs. Cependant, compte tenu de la difficulté de repérer précisément au $^{1}/_{10}$ de seconde l'occurrence des mouvements, il est souhaitable de pondérer les critères de recouvrement. En plus, que ces critères soient variables d'un type de mouvement à un autre, est un autre aspect à considérer. Par exemple, certains mouvements s'exécutent plus rapidement que d'autres, ou certains segments corporels se caractérisent par une motricité fine, à peine perceptible à l'oeil nu. La méthode de détection du mouvement doit alors s'opérer par le visionnement au ralenti de la séquence à coder. Donc, nous recommandons des critères d'écarts temporels variables par rapport à trois aspects: les frontières du recouvrement, les types de mouvement et les durées des mouvements. Les dessins ci-après illustrent des cas d'application de tels critères.

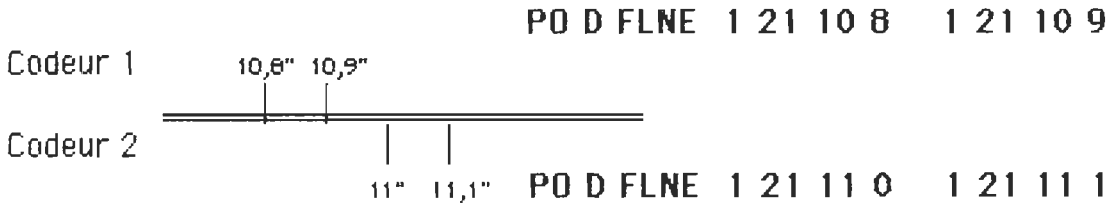


Fig. 4 -Critère de non-recouvrement acceptable pour un mouvement rapide du pouce droit

Ce type de mouvement subtil est difficilement localisable à cause de son occurrence rapide, même en observant la bande magnétoscopique à une vitesse ralentie; si son apparition s'intercale entre les unités inférieures du

chronomètre (les dixièmes de seconde), on peut supposer que deux codeurs auront de grandes chances de différer dans leur positionnement temporel. La tolérance d'un écart de l'ordre du $^{1}/_{10}$ de seconde serait acceptable dans ce cas. Par contre, pour un mouvement plus grossier, avec lequel la difficulté de repérage ressortirait aussi la vitesse d'exécution, le critère d'écart pourrait être augmenté à $^{3}/_{10}$ de seconde. Pour ce cas-ci, prenons l'exemple d'une flexion rapide de l'avant-bras, laquelle implique un plus long segment corporel que le pouce et, par conséquent, une plus longue durée d'occurrence (voir figure * 5). Ce même critère serait toutefois inadmissible pour une exécution plus lente du mouvement de flexion de l'avant-bras.

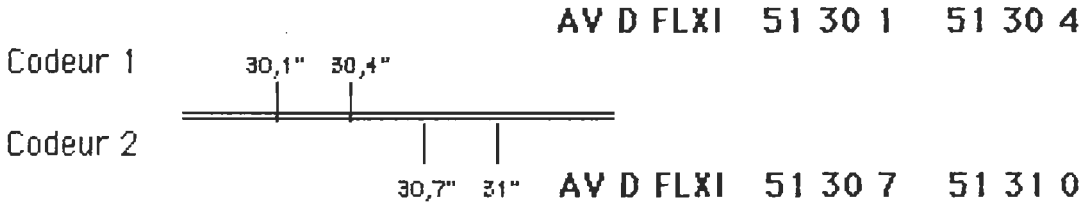
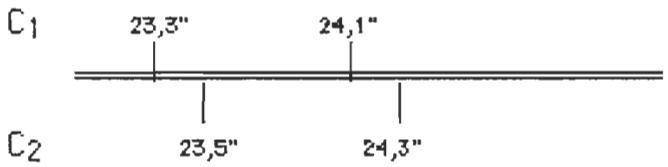


Fig. 5 - Critère de non-recouvrement acceptable pour un mouvement rapide de l'avant-bras.

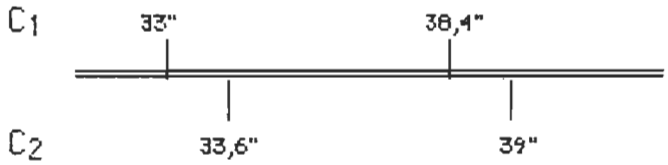
Dans les exemples (voir figure * 6) où il y a un recouvrement temporel des classements, nous constatons qu'un critère défini par le pourcentage de recouvrement est plus flexible aux variations dans les durées des mouvements qu'un critère fixe de 1, 2 ou 3, applicable aux débuts et fins de mouvements. Prenons un critère 2, soit $^{2}/_{10}$ de seconde: seuls les classements du premier exemple concorderont, alors que le deuxième exemple montre un même pourcentage de recouvrement. L'ajustement du critère à une autre unité (e.g.

aux secondes plutôt qu'aux dixièmes de seconde) est une opération fort complexe puisqu'elle nécessite une analyse à la pièce des mouvements, lesquels peuvent en plus différer dans leur temps d'exécution d'un individu à un autre.



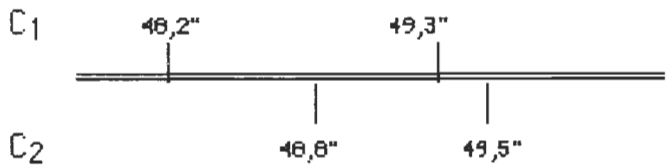
Recouvrement = 80% de la durée observée

N.B.: Si le critère de non-recouvrement acceptable est 2 ($^{2}/_{10}$ de sec.), ces classements concordent.

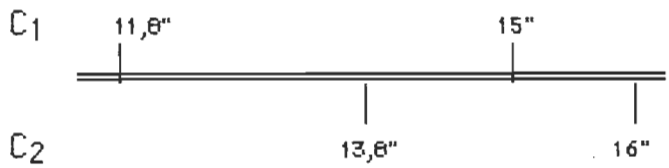


Recouvrement = 80 % de la durée observée

N.B.: Si le critère de non-recouvrement acceptable est 2 comme ci-haut, il n'y a pas de concordance.



Recouvrement = 30% de la durée observée



Recouvrement = 29% de la durée observée

Fig. 6 - Critère de recouvrement selon le pourcentage de la durée commune observée.

Plusieurs définitions de l'accord sont donc possibles dans ce modèle de «fidélité de surface»; (1) l'accord peut être basé sur le pourcentage de

mouvements différents pour lesquels deux codeurs s'entendent parfaitement, à la fois sur la catégorisation et sur les indices temporels de début et de fin de mouvement; (2) une autre forme d'accord «parfait» serait le pourcentage de mêmes types de mouvements avec des durées parfaitement concordantes; (3) des approches moins sévères pondéreraient les indices de la portion de non-recouvrement et rechercheraient un critère où l'accord atteindra son taux optimal[1]; (4) d'autres alternatives préconiseraient comme indice d'accord la proportion de la durée totale en mouvement, où deux codeurs s'entendent sur les mêmes types de mouvements, ou encore le pourcentage d'accord pour chaque type de mouvement, ou le nombre de mouvements concordants pour chaque partie du corps, ou enfin, l'addition de ces coefficients pour établir une fidélité globale. Rappelons cependant que la «fidélité de surface» ne rencontre pas les objectifs[2] du Somac puisque les interprétations que l'on veut tirer des résultats d'observation concernent surtout la quantité de mouvement observée chez un sujet. Par conséquent, l'analyse de fidélité doit inclure les taux d'amplitude observés. Ce troisième aspect de la fidélité est développé ci-après.

A présent, considérons les options pour apprécier la concordance entre deux codeurs sur la dimension spatiale des déplacements, telle qu'enregistrée

---

[1] On parle de taux optimal parce qu'il est possible de diminuer le taux d'accord même en relaxant les critères de recouvrement temporel.

[2] Le Somac a été conçu pour étudier l'organisation gestuelle d'un individu en entrevue psychologique. Ainsi, l'application prévue a trait à la typologie du mouvement chez un individu en fonction des variables suivantes: taux d'activité, fréquence et durée des mouvements, symétrie ou alternance des mouvements, dominance latérale, complexité du comportement gestuel, types de mouvements et de postures dominants, direction du mouvement, etc.

par le Somac. Cette dimension constitue le troisième critère pour accéder à une coïncidence entre deux classements lorsqu'il s'agit des mouvements nécessitant un décodage des positions spatiales adoptées par les parties du corps, c'est-à-dire de l'amplitude ou de la direction de leurs déplacements. Deux types d'enregistrement du déplacement sont donc annoncés: le déplacement codé nominalement et le déplacement digitalisé sur l'écran.

Ainsi, la codification des mouvements nominaux impliquant une description de la direction du mouvement représente une forme nominale de l'enregistrement du déplacement d'un segment corporel. Toutefois, la comparabilité de ces codes suppose l'inclusion d'autres critères entraînés par l'observation de la dimension spatiale; ils ne peuvent donc être considérés comme les codes des simples mouvements nominaux. D'un point de vue logique, il est possible d'exiger que l'accord sur les codes des mouvements nominaux avec direction soit obtenu de façon rigoureuse, c'est-à-dire avec une coïncidence exacte des catégories nominales. Cependant, dans certains cas, cette comparaison stricte passe sous silence un élément d'accord véritable; un jugement sévère, comme la non-coïncidence, s'éloigne davantage du résultat obtenu, c'est-à-dire une correspondance réelle dans les réalités observées.

Nous proposons donc ici une deuxième analyse du désaccord, ou de la non-correspondance, consistant à considérer la coïncidence dans l'orientation seule du déplacement. Par exemple, la codification d'un mouvement de l'épaule, passant de la catégorie «élévation» à la catégorie «neutre», serait considérée en accord avec la codification du même mouvement passant de la catégorie

«élévation» à la catégorie «abaissement»; la raison est que les deux codifications donnent une même direction au mouvement de l'épaule. Ce principe d'analyse est appelé le «télescopage» des codes et il a déjà été utilisé pour la fidélisation des données observationnelles. Il peut être appliqué dans le but de déterminer le niveau d'accord réel après qu'une première comparaison ait été effectuée.

La deuxième forme d'enregistrement du déplacement avec le système Somac, s'effectue par le positionnement des coordonnées cartésiennes correspondant à deux points d'articulation pour chaque segment corporel (voir figure # 3). Avec ce mode d'enregistrement, il est intéressant d'envisager un premier type de fidélité basé sur la coïncidence de ces points entre deux codeurs, au début et à la fin d'un mouvement. Cependant, étant donné la multitude de points concentrés dans la zone d'une articulation, la probabilité d'obtenir des coïncidences parfaites est faible[1]. Par exemple, un codeur peut digitaliser son point plus au centre de l'articulation à l'épaule qu'un autre codeur; ainsi, nous observerons des décalages de quelques unités sans que cela soit attribuable à une différence réelle entre les observations. Une première option pour pondérer ces écarts serait d'agrandir artificiellement un point de digitalisation à une zone circulaire de rayon défini; alors, la coïncidence des points digitalisés par deux codeurs consisterait en un croisement de deux zones (voir figure # 7).

---

[1] A la limite, avec une grille numérique suffisamment fine, la coïncidence parfaite a une probabilité nulle.

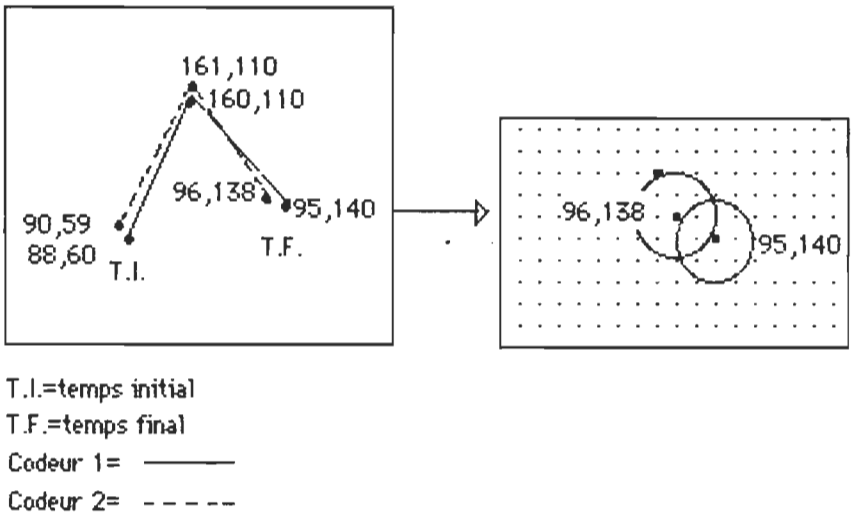


Fig. 7 - Accord défini par croisement de zones de digitalisation.

Pour établir la concordance des points digitalisés pour un même mouvement, nous avons quatre paires de coordonnées à comparer entre les codeurs, c'est-à-dire deux paires pour le début du mouvement et deux paires pour la fin du mouvement. Par exemple, comme le montre la figure * 7, le codeur * 1 a digitalisé 160,110 et 88, 60 au temps initial et 160,110 et 95,140 au temps final; le codeur * 2 a 161,110 et 90, 59, 161 110 et 96, 138. Dans ce cas, avec notre critère d'extension des coordonnées cartésiennes délimitant deux points autour du point digitalisé, nous constatons qu'il y a concordance aux quatre points entre les codeurs.

Une autre option consisterait à mesurer la surface de déplacement encourue entre le début et la fin d'un mouvement. Deux possibilités sont à envisager: (1) la surface contenue entre les quatre points digitalisés pour un mouvement (2 au début et 2 à la fin), et avec lesquels on a tracé des droites,

est calculée selon sa superficie et comparée d'un codeur à un autre (voir figure * 8);

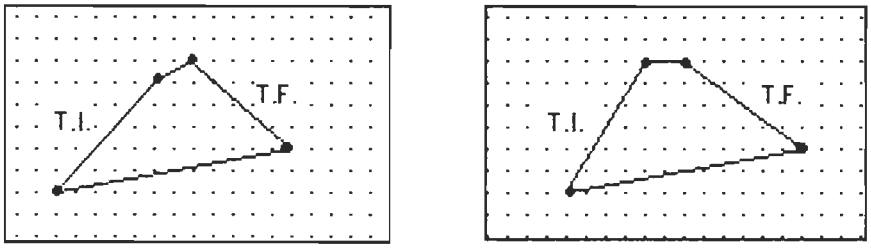


Fig. 8 - Accord défini par correspondance des surfaces de déplacement.

(2) la comparaison du déplacement effectué dans un mouvement s'établit au niveau de l'angle formé entre les deux premiers points et les deux derniers points digitalisés (voir figure * 9).

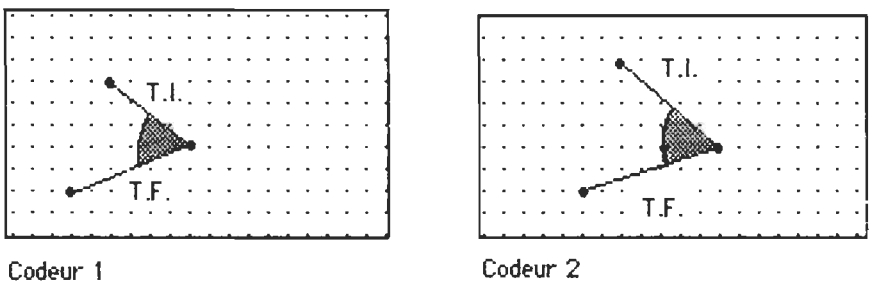
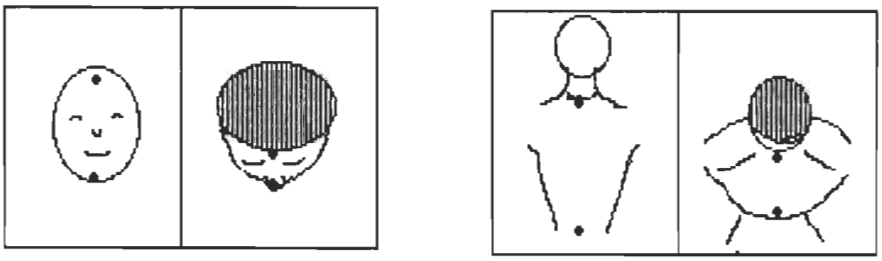


Fig. 9 - Accord défini par correspondance des angles de déplacement.

Cependant, ces deux options ne conviennent pas aux mouvements exécutés dans un plan sagittal et visionnés de face. Dans ces situations, c'est la distance entre les deux points digitalisés qui augmente ou diminue selon le type de mouvement. Ces situations se posent pour les flexions/extensions de la tête, du tronc, du bras, de l'avant-bras et de la jambe exécutées dans un plan

sagittal, c'est-à-dire de façon parfaitement perpendiculaire à la prise de vue. Le croquis #1 de la figure # 10 montre la différence entre les points de digitalisation pour une position droite de la tête et pour une flexion sagittale de la tête. Le croquis #2 illustre la même situation dans le cas d'une flexion du tronc.



1. Position droite et flexion sagittale de la tête

2. Position droite et flexion sagittale du tronc

(•)= points de digitalisation

Fig. 10 - Exemples de flexions sagittales.

Ici, la correspondance entre la longueur des droites marquées entre les points digitalisés par les codeurs, au début ainsi qu'à la fin du mouvement, indiquerait la proportion d'accord sur l'amplitude du mouvement exécuté dans un plan sagittal (voir figure # 11). Les longueurs de segments digitalisés, représentées par des droites, peuvent être comparées par le nombre de points cartésiens qu'elles couvrent.

Cette définition de la concordance comporte l'avantage d'être applicable à tous les types de mouvements. On peut comparer les droites tracées par les digitalisations de deux codeurs, de deux manières: (1) en mesurant les longueurs des segments digitalisés une à une, c'est-à-dire d'abord la longueur

du tracé digitalisé au temps initial, ensuite celle du tracé du temps final; (2) en mesurant les longueurs des segments digitalisés deux à deux, c'est-à-dire celle de la digitalisation du début et celle de la digitalisation de la fin prises comme un tout.

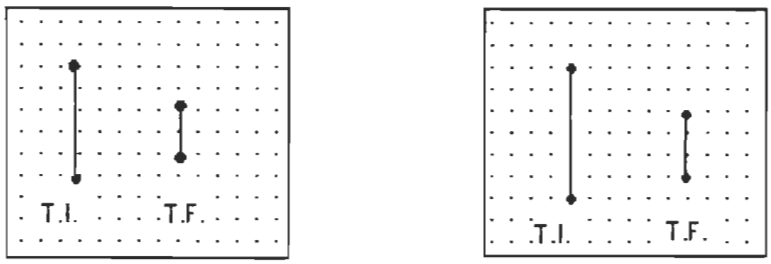


Fig. 11 - Accord défini par correspondance des longueurs de déplacement dans un plan sagittal.

Dans le cas des mouvements s'exécutant dans un plan sagittal, la première façon convient mieux puisqu'elle facilite l'accord; il est plus simple pour deux codeurs de s'entendre sur un seul segment à la fois. Cependant, si le chercheur s'intéresse en plus à l'ampleur du déplacement, la différence entre les deux longueurs de segments digitalisés deviendrait l'unité de comparaison pour l'accord. Et cette façon de mesurer l'ampleur constitue une troisième option s'ajoutant à celle de la mesure de surface du déplacement et à celle de l'angle de déplacement.

La deuxième méthode de comparaison des droites, dont il est fait mention plus haut, serait une troisième option à envisager pour tous les autres mouvements dont l'exécution s'effectue dans un plan autre que sagittal. On comparerait alors des paires de lignes en les superposant et en déterminant

leur degré de chevauchement. La superposition parfaite étant moins probable, des critères de tolérance du non-chevauchement seraient à définir pour avoir une représentation du degré de correspondance. Les croquis de la figure # 12 illustrent différents modèles de coïncidence de localisation des droites entre deux points digitalisés.

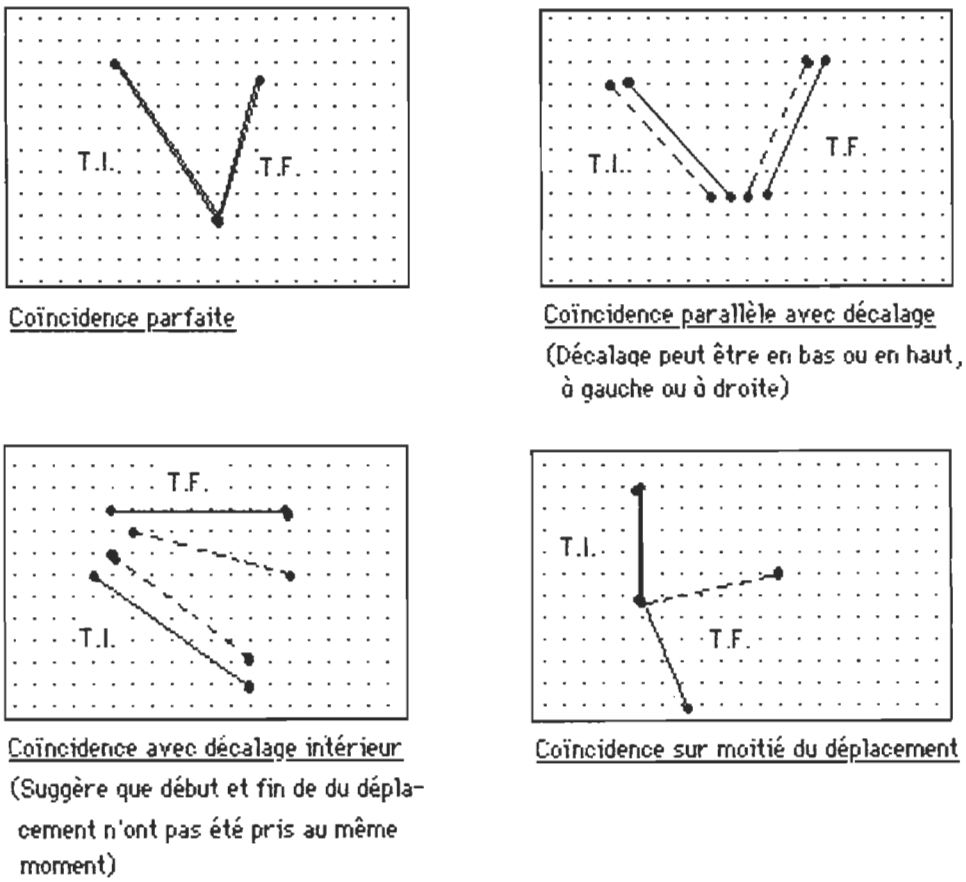


Fig. 12 - Différents types de coïncidence pour des paires de segments digitalisés par deux codeurs.

Nous avons donc présenté cinq façons d'envisager la «fidélité de déplacement»: pour les mouvements nominaux, la coïncidence des directions;

pour les mouvements normaux, la coïncidence des points de digitalisation, la correspondance des surface de déplacement, ou la correspondance des angles d'ouverture et de fermeture des mouvements, ainsi que la superposition des lignes tracées entre deux points digitalisés pour un segment corporel. Comme nous l'avons expliqué, cette cinquième option peut être abordée différemment selon le plan dans lequel s'effectue le mouvement et selon l'interprétation de la dimension «déplacement».

Toutefois, quelle que soit la méthode employée dans la fidélisation du déplacement, les différentes situations de coïncidence qui ont été illustrées nous montrent que la correspondance des localisations spatiales entre deux codeurs est en grande partie déterminée par la correspondance de leurs localisations temporelles pour un même mouvement. Cette constatation nous amène à supposer qu'il y a intérêt à produire plusieurs mesures d'accord se basant sur plusieurs niveaux d'accord, lorsque l'enregistrement observationnel comporte plusieurs dimensions simultanées. Nous introduisons ici une autre approche d'appréciation du degré d'accord comportant une dimension additionnelle d'évaluation de l'accord, soit l'exploration de la composition de l'accord. En d'autres mots, il ne suffit plus de prendre la mesure d'accord dans une acception superficielle, mais bien d'en exploiter les constituantes pour obtenir une série de facettes ou de dimensions internes qui la décrivent plus objectivement.

Pour démontrer sans trop d'élaboration notre proposition, référons-nous à la fidélité de surface utilisant les modèles de recouvrements temporels selon

les critères de pondération sélectionnés. Nous savons qu'à chaque distinction du critère de recouvrement, nous obtiendrons des proportions de coïncidence différentes; par exemple, des exigences de 30, 40, 50, 60, 70 et 80% de recouvrement nous fourniront un éventail des niveaux d'accord propices à l'illustration des différentes compositions de l'accord. Ce modèle d'analyse de l'accord s'applique tout aussi bien dans le cas de la fidélité de déplacement, où d'autres critères que le recouvrement permettront l'étalement des différents niveaux d'accords.

Nous terminons cette partie en présentant un résumé des modèles d'accord considérés. En premier, nous avons décrit la «fidélité de surface» comme l'appréciation des correspondances pour les mouvements nominaux, avec ou sans description de direction du mouvement. Ces correspondances ont été définies par la coïncidence sur:

A- La description des mouvements (partie du corps, latéralité, type de mouvement, direction)

B- La localisation temporelle, avec tolérance nulle ou positive

ou

A'- Même que A

B'- La durée des mouvements, avec tolérance nulle ou positive

En deuxième lieu, nous avons décrit la «fidélité de déplacement» comme l'appréciation des correspondances pour les mouvements normaux, de même que pour les mouvements nominaux indiquant la direction. Ces correspondances ont

été définies par la coïncidence sur:

A"- Même que A

B"- Même que B ou que B'

C- La localisation spatiale, avec tolérance nulle ou positive, selon cinq possibilités:

C_1- Points de digitalisation avec zones de recouvrement

C_2- Surfaces de déplacement entre (4) coordonnées

C_3- Angles d'amplitude des déplacements

C_4- Lignes droites représentant un segment corporel

C_5- Catégories nominales de direction

Fidélité avec une matrice de classements pour une séance de codage

Dans la section précédente, nous avons décrit plusieurs options pour l'opérationnalisation de la fidélité des données du Somac. Les diverses mesures d'accord qui ont été suggérées s'implantent dans un processus de fidélisation que nous pourrions qualifier de «à la pièce». Dans cette partie, nous désirons ébaucher un modèle de fidélisation plus global parce qu'il peut contenir à la fois toutes les dimensions faisant l'objet d'observations dans le Somac. L'avantage premier dans la conception d'une matrice des classements est donc de présenter une description cinématique des séquences de mouvements observées chez un sujet. En plus, cette matrice donnerait

l'information complète du codage effectué par un codeur lors d'une séance donnée. Bien sûr, la difficulté de concevoir des programmes informatiques, enregistrant les données de classements dans un tel tableau, est de taille et nous écartons ces aspects techniques de notre discussion. Notre objectif s'inscrit dans une démarche d'exploration des multiples possibilités de fidélisation avec un instrument tel le Somac.

Ainsi, une matrice des classements d'un codeur, analysée en superposition avec une autre matrice du même codeur, ou d'un autre codeur, pour la même séance de codage, fournirait simultanément des repères, ou des indications, sur tous les sites d'accord et de désaccord des observations multidimensionnelles. Nous voici devant une mesure d'accord continue et multidimensionnelle, se caractérisant en plus par sa représentation visuelle de la concordance intra-codeur, ou inter-codeurs. A ce moment-ci, il est justifiable de présenter concrètement notre esquisse d'une telle matrice (voir tableau * 7).

La matrice du tableau *7 présente des séries de codes pour un seul codeur. Nous avons reproduit seulement les parties du corps faisant l'objet d'une digitalisation. Les types de mouvements sont aussi indiqués en-dessous des parties du corps, ainsi que les symboles de latéralité (Droite/gauche). Les coordonnées du corps initial, enregistrées en début de séance, sont inscrites sur la première ligne de la matrice; les inscriptions de la dernière ligne représentent les coordonnées du corps final. De plus, la matrice enregistre des inscriptions de codes pour toute la durée d'un mouvement; rien ne s'inscrit s'il

s'agit d'une position neutre, ou s'il n'y a pas de mouvement.

Pour établir la fidélité à partir d'une telle matrice, il suffirait qu'une opération de juxtaposition des séries de codes de deux codeurs soit effectuée pour chaque partie du corps et chaque catégorie de mouvement; les paires de codes juxtaposés provenant de deux codeurs différents, ou du même codeur pour la première et la deuxième codification d'une séance, pourraient être transférées sur une autre matrice pour illustrer le résultat de concordance et de discordance. Donc dans un premier temps, nous aurions une matrice des classements par codeur individuel, et dans un deuxième temps, nous obtiendrions une matrice-intersection, permettant de comparer d'un seul coup tous les classements pour chaque paire de codeurs.

Il est de plus intéressant de constater que la matrice aide à dégager des zones temporelles d'activités et à localiser les parties du corps en mouvement; plusieurs autres variables peuvent ainsi être analysées pour étudier l'organisation des mouvements.

Tableau 7

Matrice des données des classements sur les 13 dimensions digitalisées

Codeur: *_____ 30" hre. min. sec. 1/10 sec.

Sujet: *_____ Longueur: _____ Début: ___ ___ ___ ___

Séance: *_____ Fin: ___ ___ ___ ___

	Tête			Tronc		Bras				Avant-bras				Cuisses		Jambes				# de mvts.
	F/E	FL	R	F/E	FL	F/E		Ad/Ab		R-Ad In/Ex	R-Ab In/Ex	F/E		F/E		F/E		R		
Temps						D	G	D	G	D	G	D	G	D	G	D	G	D	G	
T.I>0046300	F	D		F			F	Ad	Ad	Ri	Re		F	F		F	F		R	
0046301	F	D		F		E														4
0046302	F	D		F		E														4
0046303	F	D		F		E														4
0046304	F	D		F		E													R	5
0046305	F	D		F		F									F				R	6
0046306	E			E		F	E	Ab					F		F		E			8
0046307	E			E		F	E	Ab		Ri	Ri		F	E	F		E			11
0046308	E			E		F	E	Ab		Ri	Ri		F	E			E			10
0046309				E		F	E	Ab		Ri	Ri		F	E			E			9
0046310											Ri			E						2
0046311			D																	1
0046312			D																	1
0046313			D																	1
0046314																				0
0046315			G																	1
0046316			G												F		E			3
0046317			G												F		E			3
0046318			G												E		E		R	4
0046319											Ri			F	E		F		R	5
0046320	E										Ri			F	E		F		R	6
0046321	E								Ad		Ri			F	F					5
0046322	E								Ad						F		E			4
0046323					D				Ad						F		E			4
0046324					D			Ab	Ad	Re			F		F		E			7
0046325					D			Ab		Re			F				E			5
0046326					D			Ab		Re							E			4
↓																				
T.F>0047000				F				Ab	Ab			Ri	Ri	F	F	F	F		R	

Une approche géométrique de la validité des données

Nous avons discuté au chapitre deux de la quasi-absence d'études approfondies sur la validité des données d'observation naturelle. Les données recueillies avec l'instrument Somac, de par leur caractère multidimensionnel, sont une occasion de reconsidérer les procédures d'analyse de validité jusqu'ici évitées. Les données d'observation naturelle sont plus souvent déclarées intrinsèquement valides, sous prétexte de leur valeur jugée objective. Toutefois, nous ne pouvons nier le caractère interprétatif d'une observation, même avec l'utilisation d'une grille systématique.

Par conséquent, avec le cas du Somac, nous voulons poser la question suivante: "Est-ce que les classements effectués par notre système d'enregistrement peuvent être considérées comme des transcriptions adéquates des faits observés, soit les mouvements?"

En fait, l'étude du mouvement par le Somac contribue à réduire la signification de la gestuelle humaine à de simples mouvements articulaires; elle permet toutefois de conserver une certaine précision temporelle par rapport à leurs séquences d'occurrence. L'aspect psychologique de la gestuelle étant écarté, il est donc possible de reproduire le fait «mouvement» de façon plus objective, donc moins abstraite. Alors, tenter une reconstruction de la séquence observée constitue une opération assez facilement concevable. La reconstruction dont nous posons la possibilité pour les données du Somac concerne les déplacements spatiaux des différents segments corporels. Nous

désignons plus spécifiquement les déplacements enregistrés par des points digitalisés aux articulations parce que les autres ne font pas l'objet d'une observation aussi objective.

La méthode technique de reconstruction du mouvement que nous proposons consiste à introduire, par les points de digitalisation, un schéma squelettique du genre «bonhomme-allumette», lequel se superposerait au sujet observé sur l'image vidéo (voir figure # 3). Au fur et à mesure que des enregistrements de mouvements digitalisés s'effectueraient, le «bonhomme-allumette» modifierait ses positions antérieures. L'observateur aurait constamment accès à un médium de comparaison; chaque déplacement d'un segment corporel serait facilement observable à cause du prototype toujours à la vue. Il s'agit ici d'un modèle de validité continue puisque chaque déplacement enregistré se traduit par une nouvelle position du segment (ou des segments) concerné. C'est aussi une forme de validité s'établissant sur chaque unité classée, c'est-à-dire chaque segment digitalisé; elle permet donc de repérer facilement le moment où un déplacement de segment a été omis ou ajouté.

En outre, cette reconstruction schématisée des déplacements corporels inclut la possibilité de vérifier une autre forme de la validité instrumentale, soit la validité «totale», ou globale, pour une séquence de mouvements, ou à un autre niveau, pour l'ensemble d'une séance de codification délimitée dans un temps précis. Nous abordons maintenant le postulat de la cohérence des transformations spatiales du corps, selon lequel la sommation du déplacement

continu, ajoutée aux positions initiales des segments corporels du début d'une séquence de mouvement (ou d'un ensemble de séquences), correspond à leurs positions finales de la fin de la séquence, ou de la séance. Cet aspect de la validité des mesures de déplacement constitue le propre de notre argument géométrique.

En d'autres mots, nous pouvons résumer le principe de l'approche géométrique pour la validité comme suit: Prenant pour acquis que les transformations spatiales du corps enregistrées avec le Somac représentent une description objective et exhaustive des mouvements, nous obtenons un canevas illustrant le déroulement des mouvements du corps réel. Aussi, cette série de découpages montre un enchaînement de modifications spatiales géométriquement cohérentes. Nous aboutissons ainsi au principe que le «tout» (résultat final) égale la somme de ses «parties» (positions initiales + transformations).

Nous avons jusqu'ici discuté de la validité des données du Somac sous l'angle de l'élément de déplacement, lequel tient compte de la différence entre une mesure initiale et une mesure subséquente, ou finale. A présent, nous voulons enrichir notre point de vue géométrique par la simple mention que le type d'enregistrement des mouvements produit avec le Somac offre une seconde possibilité, soit de considérer la validité par l'élément de surface. Cela consisterait à obtenir une mesure de la partie du plan cartésien balayé par le mouvement et, comme pour le déplacement, à l'ajouter à la mesure d'une position initiale. La mesure de la position finale d'une séquence, ou d'une

séance, devrait correspondre au résultat de cette sommation continue. L'élément de surface peut être pris pour le corps entier, c'est-à-dire tous les angles formés aux points d'articulation entre les segments corporels, de même que pour une seule partie du corps à la fois.

Pour terminer, mentionnons que les mouvements classés seulement nominalement ne peuvent faire l'objet d'une étude de validité géométrique. La seule forme de validité praticable sur ces mouvements nominaux s'effectue par la méthode de reconstruction de Frey et Pool (1976), présentée au chapitre deux et déjà utilisée dans le cas du Somac par Déziel (1985). Ce principe de reconstruction des positions du corps débouche sur une nouvelle observation, laquelle sera comparée à l'observation initiale. Cependant, cette méthode n'apporte pas un argument solide de validité puisque la procédure de reproduction des cotes initiales comporte possiblement la même inférence.

Résultats des évaluations partielles des données recueillies avec le Somac

Nous présenterons ci-après des résultats d'analyses partielles, tirées d'une vaste banque de données obtenues avec l'instrument Somac. La mise à l'essai du Somac a comporté l'utilisation complète de toutes les catégories du Somac dans un contexte d'observation «in vivo», impliquant 13 codeurs et 19 sujets différents. Le déroulement de l'expérimentation s'est effectué principalement selon les critères suivants: (1) obtenir des mesures d'accord à différents niveaux d'habileté des codeurs; (2) répartir les codeurs en quatre équipes pour que les attributions de segments à observer s'opèrent selon un

plan en carré latin; (3) organiser le matériel d'observation pour que les segments à observer, d'égale durée, soient présentés dans un ordre différent à chaque équipe de codeurs et qu'ils soient codifiés selon une procédure uniforme; (4) obtenir des mesures d'accord où les codifications des codeurs sont comparées à un critère «de vérité», c'est-à-dire des cotes standards fournies par l'expérimentateur; (5) obtenir des mesures d'accord intra et inter-codeurs, de même que intra et inter-équipes, à chaque phase de l'expérimentation; (6) insérer une période de codification «sans contrainte», c'est-à-dire où tous les codeurs ont à codifier le même matériel d'observation, lequel est diversifié à la fois au niveau des sujets et au niveau de la complexité des déplacements articulaires.

Nos évaluations ont porté plus spécifiquement sur les données provenant de la période dite sans contrainte, où quatre codeurs ont codifié sept segments d'une durée de 30 secondes chacun. Ces sept segments avaient été prélevés des enregistrements magnétoscopiques de sept sujets à des moments différents de la séquence enregistrée. Les codeurs avaient à enregistrer les codes pour toutes les parties du corps où ils observaient des modifications. La procédure de codification s'est déroulée selon les étapes que nous avons décrites au début du chapitre III en présentant le système Somac, ainsi qu'à la section "Fidélité sur la quantité de mouvement observée"; ainsi, l'inscription des mnémoniques pour la partie du corps désignée, pour la latéralité, pour le type de mouvement, l'identification temporelle des moments de début et de fin de mouvement, et lorsque nécessaire, le repérage de l'amplitude du mouvement se sont effectués

à l'aide d'un système d'encodage informatisé.

Nous avons cependant sélectionné deux parties du corps[1] seulement pour nos analyses, soit le bras et l'avant-bras. L'observation du bras comporte la manipulation de huit catégories, soit quatre catégories par latéralité, et l'avant-bras regroupe 24 catérogies dont 12 pour chaque côté. Avant de procéder à la description de nos résultats, nous tenons à donner quelques spécifications sur le programme d'informatisation des données recueillies avec le Somac.

Les fichiers contenant les enregistrements des codeurs effectués sur un micro-ordinateur Apple II Plus, ont été transférés sur l'ordinateur central de l'Université du Québec à Trois-Rivières. L'exécution de cette opération a nécessité l'élaboration d'un programme informatique impliquant la constitution d'une multitude d'autres fichiers dont l'essence est de regrouper les données sous une structure propre aux analyses statistiques. Enfin, la conception d'un long programme en langage Fortran[2] a permis d'opérationnaliser la comparaison des fichiers de données pour établir leur niveau de concordance. Cette étude n'a pu cependant être appliquée qu'aux éléments des dimensions suivantes: les mnémoniques servant à préciser l'occurrence des mouvements et les enregistrements temporels de début et de fin de mouvement. L'étude de la

---

[1] Ces deux mêmes parties du corps ont déjà fait l'objet d'une étude de fidélité avec Déziel (1985), alors que le Somac inaugurait une procédure de codification multidimensionnelle mais avec une méthode d'encodage tout à fait différente.

[2] Nous remercions notre directeur de recherche, M. Louis Laurencelle, qui a conçu les algorithmes et le programme, le codage Fortran ayant été essentiellement réalisé par M. Yves Proulx.

concordance au niveau de la digitalisation des parties du corps en mouvement requiert l'innovation d'un programme informatique prenant en considération la nature géométrique du déplacement enregistré à partir d'une surface bididimensionnelle, soit sur la matrice de points d'un écran de télévision.

La procédure de calcul du degré d'accord inter-codeurs, avec le programme actuel d'analyse, comporte les considérations suivantes:

- pour chaque partie du corps de tous les sujets désignés, l'ordinateur sort une matrice de proportions représentant le résultat des comparaisons pour chacune des paires de codeurs à chacune des catégories;

- le chiffre inscrit au numérateur d'une proportion indique le nombre d'accords entre deux codeurs pour une catégorie spécifique;

- le chiffre inscrit au dénominateur représente le nombre total de comparaisons basées sur les inscriptions de deux fichiers de codeurs différents, pour une même catégorie; à lui seul, ce chiffre ne nous renseigne pas sur le nombre de désaccords puisque l'ordinateur ajoute une unité au dénominateur à chaque nouvelle inscription qu'il tente de comparer; en d'autres mots, cela signifie qu'au fur at à mesure que l'ordinateur lit un nouveau mnémonique dans l'un ou l'autre fichier, il tente de repérer le mnémonique correspondant chez l'autre codeur; peu importe s'il le retrouve ou si les mnémoniques coïncident, l'ordinateur enregistre une nouvelle valeur au dénominateur; le nombre de désaccords peut donc être calculé en soustrayant le numérateur du dénominateur, sans toutefois nous indiquer la nature ou la provenance de ces désaccords;

- la définition d'un accord a trait à deux aspects: la coïncidence des mnémoniques et la coïncidence des temps initiaux et finals; en ce qui concerne les inscriptions temporelles, le programme permet de calculer des indices d'accord basés sur des critères de coïncidence différents par l'addition ou la soustraction d'unités de temps constantes aux inscriptions de l'un ou l'autre codeur.

Les résultats que nous avons regroupés dans les tableaux #8, 9, 10 et 11 ont été calculés manuellement à partir des proportions affichées aux différentes matrices des sorties de l'ordinateur. Nous les rapportons aussi en nombres entiers pour apporter une description plus nuancée des pourcentages d'accord. Nous donnons des pourcentages d'accord basés sur deux approches: une approche d'accord parfait, où les temps doivent coïncider parfaitement, et une approche avec pondération, où une tolérance de ± 2 unités de temps (i.e. 2/10 de seconde) est admise pour les inscriptions temporelles de débuts et de fins de mouvement.

Le tableau 8 nous présente des indices d'accord par catégorie pour la partie du corps "bras". Nous pouvons constater que les indices sont plus élevés avec une pondération du critère d'accord. Aussi, sauf pour une situation, les indices d'accord sont plus faibles pour les catégories de la latéralité "gauche"; les deux catégories ayant le plus faible indice se retrouvent dans cette situation. Il est de plus à remarquer que les inscriptions sont plus nombreuses pour la latéralité "droite". Enfin, l'écart maximal entre les inscriptions compilées aux dénominateurs des proportions pour chaque catégorie est

représenté par la plus petite et la plus grande inscription obtenues par une paire de codeurs, ou par plusieurs paires de codeurs. La grandeur de cet écart ne montre pas une influence particulière sur la valeur des indices d'accord.

Tableau 8

Fidélité par catégorie: Résultats d'accord inter-codeurs, avec et sans critère de tolérance, pour chacune des (8) catégories de la partie du corps "BRAS"

Catégories		% d'accord moyen pour les (6) paires de codeurs		Inscriptions aux dénominateurs		
		sans critère	avec critère ± 2	Min.*	Max.*	Total
Flexion	D**	.0517	.1987	6	7	40
	G**	.0278	.2945	4	6	32
Extension	D	.0352	.1617	7	14	61
	G	.0208	.0833	4	8	37
Adduction	D	.0625	.1208	6	10	48
	G	-0-	.0238	3	7	30
Abduction	D	.1173	.202	5	9	44
	G	-0-	.125	1	4	17
% d'accord total		**.0394**	**.1512**			

* Min. et Max.= inscriptions minimales et maximales aux dénominateurs

** D=droite; G=gauche

Le tableau 9 affiche des indices d'accord par catégorie pour la partie du

corps "avant-bras". Nous constatons la même tendance à la hausse avec les indices basés sur un critère de tolérance. Cependant, on ne peut observer dans ce cas-ci une distinction constante entre les indices des catégories de l'avant-bras droit et ceux de l'avant-bras gauche. Toutefois, les six catégories totalisant un nombre d'inscriptions supérieur à 100 affichent les plus hauts indices d'accord. Enfin, il apparaît que les catégories "rotation en abduction interne et externe" montrent les plus faibles indices, alors que les catégories "flexion" et "extension" montrent les plus élevés.

Le tableau 10 donne les résultats d'accord pour chaque paire de codeurs dans les deux parties du corps "bras" et "avant-bras". Les codeurs # 1 et 12 obtiennent le plus haut degré d'accord lorsqu'il n'y a pas de critère de tolérance, alors que ce sont les codeurs # 1 et 4 lorsqu'un critère de tolérance est appliqué. De plus, le plus grand nombre d'enregistrements, ou de cotes, a été obtenu par le codeur # 1 dans les deux parties du corps, alors que c'est le codeur # 4 qui a le plus petit nombre. L'ordre pour la quantité de cotes attribuées par les codeurs est de même nature pour les deux parties du corps. Quant à l'influence du nombre d'inscriptions aux dénominateurs, elle n'apparaît pas avoir d'effet constant sur les indices d'accord.

Le tableau 11 présente les résultats d'une comparaison des indices d'accord entre deux façons de les calculer: (1) une méthode «globale» calculant une seule proportion à partir de la somme de tous les numérateurs et dénominateurs; (2) une méthode «moyenne» consistant à calculer la moyenne de toutes les proportions obtenues à chaque catégorie et à chaque paire de

Tableau 9

Fidélité par catégorie: Résultats d'accord inter-codeurs, avec et sans critère de tolérance, pour chacune des (24) catégories de la partie du corps "AVANT-BRAS"

Catégories		% d'accord moyen pour les (6) paires de codeurs		Inscriptions aux dénominateurs		
		sans critère	avec critère ± 2	Min.*	Max.*	Total
Pronation/Neutre	D**	.0625	.1412	8	9	51
	G**	.0963	.1872	10	14	73
Pronation/Supination	D	.0417	.2083	2	4	17
	G	-0-	.1667	1	1	6
Neutre/Pronation	D	.0393	.196	7	10	51
	G	.0753	.1925	11	15	79
Neutre/Supination	D	.0185	.1833	7	14	66
	G	-0-	.07	8	14	70
Supination/Pronation	D	.0555	.1527	3	4	20
	G	.0833	.0833	1	2	9
Supination/Neutre	D	.0313	.1777	8	16	70
	G	.0762	.1238	9	15	71
Rotation en adduction	D	.0098	.121	4	17	76
interne	G	.0743	.2185	14	26	126
Rotation en adduction	D	.0167	.1063	9	16	71
externe	G	.0712	.1542	13	27	126
Rotation en abduction	D	.0167	.0333	13	20	85
interne	G	-0-	-0-	2	14	46
Rotation en abduction	D	.0138	.0208	16	24	93
externe	G	.0334	.0334	2	12	40
Flexion	D	.1837	.308	15	27	136
	G	.1118	.2023	19	32	160
Extension	D	.0783	.1668	16	31	150
	G	.1181	.2953	21	32	164
% d'accord total		**.0545**	**.1476**			

* Min. et Max.= inscriptions minimales et maximales aux dénominateurs

** D=droite; G=gauche

Tableau 10

Fidélité par paire de codeurs: Résultats d'accord inter-codeurs, avec et sans critère de tolérance, pour chacune des (6) paires de codeurs

Numéros des codeurs	% d'accord moyen pour toutes les catégories d'une partie du corps		Inscriptions totales aux dénominateurs	# total de cotes par codeur
	sans critère	avec critère ± 2		
Bras				
1 & 4	.0281	.2013	51	# 1 = 48
1 & 12	.0804	.1931	54	# 4 = 24
1 & 22	.0139	.1122	65	# 12 = 33
4 & 12	.0125	.1378	37	# 22 = 45
4 & 22	.0295	.0947	51	
12 & 22	.0595	.168	51	
% total	**.0394**	**.1512**		
Avant-bras				
1 & 4	.0697	.1922	336	# 1 = 297
1 & 12	.0716	.1219	346	# 4 = 181
1 & 22	.0623	.1276	348	# 12 = 215
4 & 12	.0335	.1728	235	# 22 = 216
4 & 22	.0687	.1532	285	
12 & 22	.0225	.1309	306	
% total	**.0547**	**.1498**		

codeurs. Les indices d'accord global sont légèrement plus élevés que les indices d'accord moyen. Encore ici, on peut constater que la grandeur du nombre d'inscriptions n'exerce pas d'influence sur le degré de concordance.

Tableau 11

Comparaison de la fidélité globale et de la fidélité moyenne: Résultats d'accord pour tous les codeurs et pour toutes les catégories, avec et sans critère de tolérance, selon une méthode de calcul global et selon une méthode de calcul de la moyenne des proportions par catégorie

Partie du corps	% d'accord Sans critère		Total aux numérateurs & aux dénominateurs	% d'accord Avec critère ± 2		Total aux numérateurs & aux dénominateurs
	Global	Moyen		Global	Moyen	
Bras	.0453	.0394	14/309	.1489	.1512	46/309
Avant-bras	.0717	.0546	133/1856	.1681	.1487	312/1856
TOTAL	**.0585**	**.047**		**.1585**	**.1499**	

Nous concluons cette section en apportant quelques commentaires sur les résultats partiels que nous venons de présenter. Mentionnons d'abord que les indices d'accord apparaissent nettement plus faibles que ceux que les

recherches observationnelles rapportent habituellement. Les pourcentages dans nos tableaux représentent l'accord moyen de deux observateurs pour attribuer la catégorie désignée, cet accord moyen étant évalué à travers une somme d'événements. Il s'agit donc d'un accord de «commission» seulement, non pas d'un pourcentage d'accords dérivé du tableau «2x2» et correspondant à 100 (A+D)/(A+B+C+D), où 'A' dénote les accords de commission et 'D' les accords d'omission: notre indice ci-haut ne considère pas le terme 'D', qui représente usuellement la plus grande part du pourcentage observé et qui a motivé à lui seul les "corrections pour le hasard" que l'on retrouve par exemple avec le 'k' de Cohen.

Il faut ajouter à cela que le pourcentage d'accords dans nos tableaux reflète une «fidélité par item», puisqu'il rapporte l'accord moyen par événement pour une catégorie donnée: à cela correspondrait un coefficient de fidélité classique, corrigé par la formule de Spearman-Brown. Par exemple, pour un test psychométrique de 100 items et ayant une fidélité globale (de type test-retest) de 0,90, la fidélité par item serait en moyenne de 0,08. Le faible pourcentage d'acords, rapporté dans nos tableaux 8 à 11, ne représente donc pas nécessairement une si mauvaise performance de nos observateurs avec le système Somac.

Mentionnons de plus qu'un seul niveau de critère de tolérance, soit 2/10 de seconde, a été utilisé pour évaluer la concordance; on peut supposer que la marge temporelle optimale est autre que le ± 2, et qu'une analyse systématique comportant plusieurs niveaux de critère résulterait dans l'obtention d'un

plateau en dehors duquel les indices s'abaisseraient.

Rappelons en outre que, dans le cas du Somac, la définition de la concordance inclut un critère temporel basé sur le 1/10 de seconde, lequel représente une sévère exigence pour atteindre une coïncidence établie sur une telle micro-unité d'observation, comparativement aux unités d'observation des recherches courantes reposant sur un critère temporel fixe et pré-déterminé, comme par exemple un intervalle de 10 secondes.

De plus, il est justifié d'anticiper un niveau d'accord plus faible avec des unités d'observation multidimensionnelles puisque plusieurs conditions doivent être rencontrées simultanément pour justifier une coïncidence. Dans le cas de l'observation unidimensionnelle de type «présence/absence», la concordance est facilement atteignable parce qu'étant définie par un principe de «tout ou rien».

La visualisation des fichiers de données du Somac permet de constater qu'il y a concordance dans l'identification des mouvements mais que l'écart apparaît au niveau de la précision temporelle. Ceci nous amène à supposer que le repérage temporel continu, sur une unité aussi petite que le 1/10 de seconde, contribue pour une grande part à la complexité d'effectuer une inscription et, par conséquent, diminue la probabilité d'obtenir des coïncidences parfaites pour ces enregistrements.

L'inspection des fichiers de données nous permet aussi d'identifier une différence accentuée entre le nombre d'attributions pour une même catégorie.

Les écarts d'inscriptions aux dénominateurs relatés dans les tableaux 8 et 9 ainsi que le nombre d'inscriptions enregistrées par chacun des codeurs et affiché au tableau 10, font état de ce désaccord dans la quantité de cotes attribuées par les codeurs. Nous en déduisons qu'un certain degré d'ambiguïté persiste par rapport à la reconnaissance de quelques catégories. Ce «flottement» sur la quantité d'occurrences dans une catégorie apporte des interrogations par rapport à la procédure d'entraînement des codeurs. Ainsi, une évaluation de la «fidélité par item» pour toutes les parties du corps est nécessaire pour mieux commenter les imprécisions de certaines catégories.

Finalement, nous tenons à souligner l'avantage indéniable de la procédure d'encodage des données du Somac, lequel se rattache à la possibilité d'effectuer une évaluation complète des enregistrements obtenus. Cette caractéristique du Somac marque une grande avance sur la plupart des recherches observationnelles, lesquelles n'appliquent que des évaluations sporadiques de leurs données et, dans bien des cas, sur une infime quantité de données par rapport à l'ensemble de celles recueillies pour tirer leurs conclusions sur les variables comportementales observées.

Conclusion

Notre recherche a contribué à montrer la diversité des systèmes d'enregistrement et d'échantillonnage du comportement. Par le long recensement des approches de fidélisation des données, nous avons fait ressortir que le choix d'un indice de fidélité ne faisait pas consensus entre les chercheurs. Aussi, notre constat explique que la polémique tient surtout à la difficulté de transférer les concepts psychométriques traditionnels au domaine des données d'observation.

Le problème étant mieux défini, nous avons voulu initier un mouvement empirique pour relancer les efforts de recherche vers une conceptualisation plus universelle du processus observationnel. Nous avons choisi une grille d'observation comportant des mesures multidimensionnelles et continues afin de présenter des techniques d'évaluation de l'accord entre observateurs jusqu'à maintenant peu élaborées. Les difficultés n'ont certes pas été toutes résolues; la complexité de confectionner des programmes informatiques pour les analyses statistiques en est une de taille. Néanmoins, nos suggestions par rapport aux approches de calcul de la concordance des catégorisations spatio-temporelles sont réalisables avec les outils techniques actuels. En outre, les procédés nouveaux de validation instrumentale, comme la matrice des classements et l'argument géométrique, ne sauraient être contestés quant à

leurs retombées dans plusieurs domaines de recherche tels l'ergonomie, la biomécanique, l'entrevue psychologique, etc.

L'importance d'étudier le comportement avec des instruments dont la qualité métrique a été démontrée par des méthodes pertinentes et incontestées nous apparaît tributaire de la valeur des observations obtenues dans un contexte éthologique.

Remerciements

L'auteure désire exprimer sa reconnaissance à son directeur de thèse, monsieur Louis Laurencelle, Ph. D., pour son appui constant et sa compétente assistance durant la réalisation de ce mémoire.

Des remerciements sont aussi adressés à madame Micheline Dubé, D. Ps., pour son encadrement en tant que directrice du projet de recherche sur l'utilisation de l'instrument Somac et comme co-directrice du mémoire de recherche.

L'auteure désire également souligner la précieuse contribution de monsieur Yves Proulx, M.Sc., notamment, comme concepteur du programme d'acquisition des données. De plus, l'auteure est redevable des efforts de nombreux autres collaborateurs, grâce à qui le travail de cotation a pu être effectué.

Enfin, la contribution financière du Conseil de recherche en sciences naturelles et en génie du Canada, par le biais de leur Bourse d'Etudes Supérieures, constitua un apport appréciable pour la réalisation de ce mémoire.

Références

ALTMANN, J. (1974). Observational study of behaviour: Sampling methods. Behaviour, 49, 227-267.

BAER, D. M. (1977). Reviewer's comment: Just because it's reliable doesn't mean that you can use it. Journal of applied behavior analysis, 10, 117-119.

BAKEMAN, R. (1978). Untangling streams of behavior: Sequential analyses of observation data, in G. P. Sackett, G. C. Ruppenthal & J. Gluck (Eds.): Observing behavior, Vol. II: Data collection and analysis methods (pp.63-78). Baltimore: University Park Press.

BARNETT, S. A. (1980). Ethology and Man: Science or Myth?, in S. A. Corson & E. O. Corson (Eds.): Ethology and nonverbal communication in mental health (pp.47-61). New York: Pergamon Press.

BEAUGRAND, J. P. (1982). Observation directe du comportement, in M. Robert (Ed.): Fondements et étapes de la recherche scientifique en psychologie (pp.167-218). Montréal: Chenelière et Stanké.

BELANGER, D. (1982). Mesure des phénomènes, in M. Robert (Ed.): Fondements et étapes de la recherche scientifique en psychologie (pp. 153-166). Montréal: Chenelière et Stanké.

BERK, R. A. (1979). Generalizability of behavioral observations: A Clarification of interobserver agreement and interobserver reliability. American journal of mental deficiency, 83, 460-472.

BERNIER, J. J. (1985). Théorie des tests: Principes et techniques de base. Chicoutimi, Québec: Gaëtan Morin.

BIRKIMER, J. C. & BROWN, J. H. (1979). Back to basics: Percentage agreement

measures are adequate but there are easier ways. Journal of applied behavior analysis, 12, 535-543.

BOVET, P. (1984). Observation et statistiques, in M.-P. Michiels-Philippe et coll. (Eds.): L'observation, 2è Partie (pp. 73-97). Paris: Delachaux et Niestlé.

BUNGE, M. (1984). L'observation, in M.-P. Michiels-Philippe et coll. (Eds.): L'observation, 1ère Partie (pp. 47-49). Paris: Delachaux et Niestlé.

CARO, T. M., ROPER, R., YOUNG, M., DANK, G. R. (1979). Inter-observer reliability. Behaviour, 69, 303-315.

CONDON, W. S. (1976). Une analyse de l'organisation comportementale, in J. Cosnier et A. Brossard (Ed.): La communication non verbale (pp. 31-70). Paris: Delachaux & Niestlé, 1984.

CONE, J. D. (1977). The relevance of reliability and validity for behavioral assessment. Behavior therapy, 8, 411-426.

CONE, J. D. (1982). Validity of direct observation assessment procedures. New Directions for methodology of social and behavioral science, 14, 67-79.

CONGER, A. J. (1985). Kappa reliabilities for continuous behaviors and events. Educational and psychological measurement, 45, 861-868.

CONGER, A. J. & WARD, D. G. (1984). Agreement among 2x2 agreement indices. Educational and psychological measurement, 44, 301-314.

COSNIER, J., BROSSARD, A. (1984). Communication non verbale: co-texte ou contexte?, in J. Cosnier et A. Brossard (Ed.): La communication non verbale (pp. 1-29). Paris: Delachaux & Niestlé.

DEZIEL, P. (1985). Validation d'une grille d'observation du comportement non verbal en considérant les mouvements des bras et des avant-bras. Mémoire de maîtrise inédit, Université du Québec à Trois-Rivières.

DUBE, M., PELLERIN, A., DEZIEL, P. & CHARRIER, J. (1985). The Somac: A new

approach for decoding body movement parameters. Perceptual and Motor Skills, 61, 1106.

EKMAN, P., FRIESEN, W. V. (1976). La mesure des mouvements faciaux, in J. Cosnier et A. Brossard (Ed.): La communication non verbale (pp. 101-124). Paris: Delachaux et Niestlé, 1984.

ELSTON, R. C., SCHROEDER, S. R., ROJAHN, J. (1982). Measures of observer agreement when binomial data are collected in free operant situations. Journal of behavioral assessment, 4, 299-310.

FASSNACHT, G. (1982). Observational systems, in G. Fassnacht (Ed.): Theory and practice of observing behaviour (pp. 83-114). New York Academic Press.

FASSNACHT, G. (1982). Quantifying observations, in G. Fassnacht (Ed.): Theory and practice of observing behaviour (pp. 115-136). New York Academic Press.

FASSNACHT, G. (1982). Two Perspectives on observation: The ethological and ecological, in G. Fassnacht (Ed.): Theory and practice of observing behaviour (pp. 137-180). New York Academic Press.

FLEISS, J. L. (1971). Measuring nominal scale agreement among many raters. Psychological bulletin, 76, 378-382.

FOSTER, S. L., CONE, J. D. (1980). Current issues in direct observation. Behavioral Assessment, 2, 313-338.

FREY, S., HIRSBRUNNER, H.-P., FLORIN, A.-M., DAW, W., CRAWFORD, R. (1983). Analyse intégrée du comportement non verbal et verbal dans le domaine de la communication (Adaptation française, A. Brossard, 1984), in J. Cosnier et A. Brossard (Ed.): La communication non verbale (pp. 145-227). Paris: Delachaux & Niestlé, 1984.

FREY, S. & POOL, J. (1976-2). A new approach to the analysis of visible behavior. Research reports form the Department of Psychology, University of Berne, Suisse.

HARRIS, F. C. & LAHEY, B. B. (1978). A method for combining occurrence and nonoccurrence interobserver agreement scores. Journal of applied behavior analysis, 11, 523-527.

HARRIS, F. C. & LAHEY, B. B. (1982). Recording system bias in direct observational methodology. A review and critical analysis of factors causing inaccurate coding behavior. Clinical psychology review, 2, 539-556.

HARTMANN, D. P. (1977). Considerations in the choice of interobserver reliability estimates. Journal of applied behavior analysis, 10, 103-116.

HARTMANN, D. P. (1982). Assessing the dependability of observational data. New directions for methodology of social and behavioral science, 14, 51-65.

HINDE, R. A. (1982). Ethology: Its nature and relations with other sciences. New York: Oxford University Press.

HOLLENBECK, A. R. (1978). Problems of reliability in observational research, in G. P. Sackett, G. C. Ruppenthal & J. Gluck (Eds.): Observing behavior, Vol. II: Data collection and analysis methods (pp. 79-98). Baltimore: University Park Press.

HOLM, R. A. (1978). Techniques of recording observational data, in G. P. Sackett, G. C. Ruppenthal & J. Gluck (Eds.): Observing behavior, Vol. II: Data collection and analysis methods (pp. 99-108). Baltimore: University Park Press.

HOPKINS, B. L. (1979). Proposed conventions for evaluating observer reliability: A commentary on two articles by Birkimer and Brown. Journal of applied behavior analysis, 12, 561-564.

HOUSE, A. E., HOUSE, B. J. & CAMPBELL, M. B. (1981). Measures of interobserver agreement: Calculation formulas and distribution effects. Journal of behavioral assessment, 3, 37-57.

JOHNSON, S. M. & BOLSTAD, O. D. (1973). Methodological issues in naturalistic observation: Some problems and solutions for field research, in L. A. Hamerlynck, L. C. Handy & E. J. Mash (Eds.): Behavior change, methodology, concepts and practice (pp. 7-67). Champaign, Il.: Research Press.

KAZDIN, A. E. (1977). Artifact, bias, and complexity of assessment: The ABC's of reliability. Journal of applied behavior analysis, 10, 141-150.

KELLER, H. R. (1980). Issues in the use of observational assessment. School Psychology review, 9, 21-30.

KELLY, M. B. (1977). A review of the observational data-collection and reliability procedures reported in the Journal of Applied Behavior Analysis. Journal of applied behavior analysis, 10, 97-101.

KENT, R. N., FOSTER, S. L. (1977). Direct observational procedures: Methodological issues in naturalistic settings, in A. R. Ciminero, K. S. Calhoun, H. E. Adams (Eds.): Handbook of behavioral assessment (pp. 279-328). New York: Wiley.

KRATOCHWILL, T. R. & WETZEL, R. J. (1977). Observer agreement, credibility, and judgment: Some considerations in presenting observer agreement data. Journal of applied behavior analysis, 10, 133-139.

LAURENCELLE, L. (1986). Observer le réel: Quelques questions d'intérêt méthodologique, in C. Paré, M. Lirette, M. Piéron (Eds.): Méthodologie de la recherche en enseignement de l'activité physique et sportive, Trois-Rivières: Université du Québec à Trois-Rivières.

LAURENCELLE, L. (1983). Une interprétation du pourcentage d'accords de la fidélité inter-juges. Lettres statistiques, 7, chap. 4, 12 p.. Document non-publié, Trois-Rivières: Université du Québec à Trois-Rivières.

LAURENCELLE, L., RAFMAN, S. (1981). Dyadic interaction and change patterns in child psychotherapy: Rapport méthodologique. Santé et Bien-Etre, Ottawa.

LONGABAUGH, R. (1980). The systematic observation of behavior in naturalistic settings, in H. C. Triandis (Ed.): Handbook of cross cultural

psychology (pp. 57-127). Boston, Toronto: Allyn and Bacon.

McDOWALL, J. J. (1978). Microanalysis of filmed movement: The reliability of boundary detection by observers. Environmental psychology and nonverbal behavior, 3, 77-88.

MITCHELL, S. K. (1979). Interobserver agreement, reliability, and generalizability of data collected in observational studies. Psychological bulletin, 86, 376-390.

RUNKEL, P. J. & McGRATH, J. E. (1972). Planning to gather evidence: Validity, reliability, and generalizability. Research on human behavior: A systematic guide to method (pp. 149-172). Holt, Rinehart and Winston.

SACKETT, G. P. (1978). Measurement in observational research, in G. P. Sackett, G. C. Ruppenthal, & J. Gluck (Eds.): Observing behavior, Vol. II: Data collection and analysis (pp. 25-43). Baltimore: University Park Press.

SACKETT, G. P., LANDESMAN-DWYER, S. (1982). Data analysis: Methods and problems. New directions for methodology of social and behavioral science, 14, 81-99.

SACKETT, G. P., RUPPENTHAL, G. C., GLUCK, J. (1978). Introduction: An overview of methodological and statistical problems in observational research, in G. P. Sackett, G. C. Ruppenthal, & J. Gluck (Eds.): Observing behavior, Vol. II: Data collection and analysis methods (pp. 1-14). Baltimore: University Park Press.

SCOTT, M. M., HATFIELD, J. G. (1985). Problems of analyst and observer agreement in naturalistic narrative data. Journal of educational measurement, 22, 207-218.

TOWSTOPIAT, O. (1984). A review of reliability procedures for measuring observer agreement. Contemporary educational psychology, 9, 333-352.

TRUDEL, M. & STRAYER, F. (1986). Les démarches observationelles et les questions de validité. Revue canadienne de l'étude en petite enfance, 1, 169-176.

VAUCLAIR, J. (1984). L'observation en éthologie, in M.-P. Michiels-Philippe et coll. (Eds.): L'observation (pp.123-136). Paris: Delachaux et Niestlé.

WAKEFIELD, J. A. Jr. (1980). Relationship between two expressions of reliability: Percentage agreement and Phi. Educational and psychological measurement, 40, 593-597.

WILDMAN, B. G. & ERIKSON, M. T. (1977). Methodological problems in behavioral observation, in R. Hawkins & J. D. Cone (Eds.): Behavioral assessment: New directions in clinical psychology (pp. 255-273). New York: Brunner/Mazel.

YELTON, A. R. (1979). Reliability in the context of the experiment: A commentary on two articles by Birkimer and Brown. Journal of applied behavior analysis, 12, 565-569.

YELTON, A. R., WILDMAN, B. G. & ERIKSON, M. T. (1977). A probability-based formula for claculating interobserver agreement. Journal of applied behavior analysis, 10, 127-131.